Gender discrimination in hiring across occupations: a nationally-representative vignette study

Kübler, Dorothea; Schmid, Julia; Stüber, Robert

Kübler, Dorothea; Schmid, Julia; Stüber, Robert

Article — Accepted Manuscript (Postprint)

Gender discrimination in hiring across occupations: a nationally-representative vignette study

Labour Economics

Provided in Cooperation with:
WZB Berlin Social Science Center

Suggested Citation: Kübler, Dorothea; Schmid, Julia; Stüber, Robert (2018) : Gender discrimination in hiring across occupations: a nationally-representative vignette study, Labour Economics, ISSN 0927-5371, Elsevier, Amsterdam, Vol. 55, pp. 215-229, http://dx.doi.org/10.1016/j.labeco.2018.10.002

This Version is available at:
http://hdl.handle.net/10419/213866

Published version: https://doi.org/10.1016/j.labeco.2018.10.002

Gender Discrimination in Hiring Across Occupations: A Nationally-Representative Vignette Study*

Dorothea Kübler† Julia Schmid‡ Robert Stüber§

August 24, 2018

Abstract

We investigate gender discrimination in a nationally-representative sample of German firms using a factorial survey design. Short CVs of fictitious applicants for apprenticeship positions are presented to human resource managers who are asked to evaluate the applicants. Women are evaluated worse than men on average, controlling for all attributes of the CV. This measure of discrimination is robust to differences in the variance of unobservable productivity characteristics ("Heckman critique"). Discrimination against women varies across industries and occupations. Controlling for all occupation- and firm-related variables that we observe, only the share of women in an occupation correlates with discrimination.

JEL Classification: *C99; J16; J71*

Keywords: Gender discrimination, hiring decisions, vignette study

*We would like to thank Heike Solga and Paula Protsch for their collaboration and fruitful discussions, the BIBB (Federal Institute of Vocational Education and Training) which made the study possible, and Katrin Auspurg who helped us with the vignette selection. We are grateful to Bernd Fitzenberger, Patrick Gauß, Stephan Martin, Kristina Strohmaier, and Marina Töpfer for their comments, to Thu-Ha Nguyen, Michaela Zwiebel, Sajoscha Engelhardt, and Manuela Ludwig for their assistance, and to Jennifer Rontganger for copy editing. We thank the editor and three referees for their constructive comments. Financial support from CRC TRR 190 is gratefully acknowledged. A previous version of the paper was circulated under the title "Be a Man or Become a Nurse: Comparing Gender Discrimination by Employers Across a Wide Variety of Professions."

†WZB, Reichpietschufer 50, 10785 Berlin, Germany, and Technical University Berlin; e-mail: kuebler@wzb.eu

‡DIW Berlin, Mohrenstraße 58, 10117 Berlin, Germany; e-mail: jschmid@diw.de

§Berlin Doctoral Program in Economics and Management Science (BDPEMS) and WZB, Reichpietschufer 50, 10785 Berlin, Germany; e-mail: robert.stueber@wzb.eu

1 Introduction

Some labor market outcomes of men and women, such as participation rates and earnings, have converged in recent decades. However, major gender differences in the labor market prevail (see, e.g., Blau and Kahn, 2007; Bureau of Labor Statistics, 2017; Upright, 2017). It is an open question what causes these persistent differences. While different trajectories in the labor market between men and women may be due to differences unfolding in the life course, such as asymmetries in the effects of having children (see Kleven et al., 2017), we focus on the first stage of people's careers. Since girls, on average, do at least as well in school as boys, their chances of getting a job after school should be at least as good. We thus consider the question of whether gender discrimination sets in early by studying the employment probabilities of men and women right after leaving school when applying for their first job.[1] By gender discrimination we mean that men and women who are equal with respect to all observable productivity-related characteristics are evaluated differently by employers in our survey.[2]

We designed a factorial survey based on vignettes consisting of the CVs of fictitious applicants. Factorial survey designs were recently used by economists to study discrimination in the labor market (Lahey and Oxley, 2017; Finseraas et al., 2016; Baert and De Pauw, 2014), fairness perceptions (Stephan et al., 2014) and ethical judgments (Ambuehl and Ockenfels, 2017; Ambuehl et al., 2015). The respondents of the survey, mainly firm owners and human resource managers, were asked to evaluate the fictitious applicants as if they had applied for an apprenticeship position in the firm. Thus, the managers were asked to perform a task that resembles their everyday work. Moreover, we draw the survey sample from the target population, which is likely to increase the external validity of our study (Hainmueller et al., 2015). Each respondent evaluated several fictitious applicants of the same gender. This reduces potential spillover effects between applications while still maintaining a large sample size (see Lahey and Beasley, 2018; Phillips, 2016).

Since the respondents only make hypothetical choices, the method does not allow us to observe actual hiring decisions or interview invitations. However, the vignette design can complement the use of observational data and field experiments to study discrimination for the following reasons. First, we included the vignette study in a nationally-representative survey, thereby obtaining evaluations by German firms that hire apprentices in 126 different professions. The statistical population of our study encompasses all firms of all industries and sizes in Germany. Hence, our sample is representative of a larger population than the samples usually used in vignette and field

[1] There is evidence of an existing gender pay gap at career entry and at early career stages in Germany. Kunze (2005, 2003) reports evidence of a wage differential of more than 20 percent between female and male full-time workers who have participated in the dual apprenticeship system. More recently, Behr and Theune (2015) find that starting salaries are lower for women for the entire income distribution. There is also evidence of wage differentials at the beginning of the career in other countries such as the US (Marini and Fan, 1997).

[2] This definition encompasses taste-based and statistical discrimination. Note that statistical discrimination is seen as acceptable in some contexts (young drivers have to pay more for car insurance, for example), while it is considered as unacceptable or illegal in other contexts, such as gender- or race-based hiring decisions.

studies on discrimination.[3]

Second, tying the vignettes to a nationwide survey allows us to not only estimate discrimination within a large variety of occupations and industries, but also to link the evidence from the vignettes with a rich dataset on the firms. This allows us to investigate the firm- and occupation-specific determinants of gender discrimination and to disentangle their effects. Moreover, it enables us to shed light on the nature of discrimination by testing for statistical discrimination that is due to differences in the precision of the observable productivity components (Aigner and Cain, 1977).

Third and finally, as with field experiments we are able to estimate the causal effect of being female on our outcome variable (the evaluations) and, by varying different applicant characteristics, to analyze how discrimination varies with applicant characteristics and the quality of an application. Heckman and Siegelman have critized field experiments that do not vary the quality of the applicant by pointing out that if the variance of unobserved productivity characteristics differs between men and women, taking a snapshot at one quality level of applicants (as most field experiments do) can generate a biased measure of discrimination (Heckman, 1998; Heckman and Siegelman, 1993). Neumark (2012) developed a method to recover an unbiased estimate of discrimination despite this issue, and his method has been applied in various studies on discrimination (Baert, 2015; Carlsson et al., 2014; Neumark et al., 2016; Neumark and Rich, 2016; Neumark et al., 2015). We provide an alternative way to address the critique by using a factorial survey design. Our outcome variable is non-binary and there is substantial variation in both applicant quality and employer requirements. Therefore, a difference in the variance of unobserved productivity characteristics cannot bias our estimate of discrimination.

We focus on the market for apprenticeships, which is the main entry-level labor market for young Germans below the level of tertiary education. More than 60 percent of all school-leavers start an apprenticeship. Besides the possibility to choose a profession for which training is exclusively provided in vocational schools (such as nursing, physiotherapy, or early childhood education), the only way to get training outside of universities is to participate in the apprenticeship system. Apprentices gain practical work experience in a firm and go to a vocational school where they are taught subjects such as German, math, and social studies. Hence, apprenticeships have an important educational component, which differentiates them from regular employment. As many jobs in Germany require having completed an apprenticeship, discrimination in the apprenticeship market can be the source of different competencies that lead to persistent differences in gender ratios across professions. In fact, in 2015, out of all apprentices 61.9 percent were men, and only 38.1 percent were women.

The hiring procedures for apprentices are similar to those for other employees, as applicants have to apply at the firm which will train them. As for regular employment, there is large variation

[3]Vignette studies on discrimination often use students (Correll et al., 2007; Heilman, 1984; Lahey and Oxley, 2017; Baert and De Pauw, 2014; Blommaert et al., 2014) or target groups that attend a particular event (Rosen and Jerdee, 1974).

in the length and professionalization of the recruitment process, and salaries vary substantially between professions and industries. The market for apprenticeships is competitive in that a number of applicants do not get a job in their desired profession, and many applicants remain without a job every year. At the same time, many slots are unfilled due to strong regional and occupational differences (Federal Employment Agency, 2017). Although employers recruit apprentices for the duration of the apprenticeship and there is no commitment for the employers to employ the apprentice after the training has ended, more than two thirds of the apprentices remain with the firm that trained them (see BIBB, 2017).

We find economically substantial and statistically significant discrimination against women in the evaluations of the fictitious CVs. The penalty for being female is as large as the effect of having a grade point average that is worse by one grade.[4] This estimate is robust to different variances of unobserved productivity characteristics between men and women. We find no evidence that the discrimination is caused by differences in the precision of the observable productivity components.

The amount of discrimination varies by industry. In line with prior research (e.g., Riach and Rich, 2006), it matters whether an occupation is predominantly female or male, with women being significantly less likely to be invited for an interview than men when applying for male-dominated occupations. We do not find evidence of discrimination against men in female-dominated occupations, though. There is no evidence of discrimination if the labor market is very tight. The size of the firm (as measured by the number of employees or the number of apprentices) and the degree of professionalization of the recruitment process do not matter as moderators of gender discrimination. Finally, we study whether the average salary, the typical school-leaving qualification required, and the status of a profession moderate discrimination. Prior research suggests that women experience discrimination in high-status occupations (e.g., Neumark et al., 1996). We find that women fare especially worse than men in professions with a low social status according to the status index. The educational requirement and the average wage of a profession have no systematic impact on gender differences in evaluations.

Existing studies on the role of the share of women in a profession have looked at a limited number of professions and are unable to control for the status of a profession (for example, secretaries as typically female and engineers as typically male occupations). A few studies have looked at the role of occupational status in explaining discrimination, but are again based on a few professions only and are unable to systematically control for the share of women. We aim at disentangling the different characteristics of the professions and provide the correlations between the potential moderator variables and gender discrimination. Our main finding is that, controlling for all moderators, only the share of women in a profession correlates with the difference in evaluations. All other firm- and occupation-specific variables cannot explain variation in the discrimination observed. These results are novel because they are derived from a large representative sample of firms that allows us to control for many possible moderators.

[4]German school grades range from 1 (best) to 6 (worst) where 4 is the passing grade.

2 Literature and research questions

The study relates to the literature on gender discrimination in the labor market as well as to methodological contributions regarding the identification of discrimination. We discuss these two strands of the literature in turn.

Gender discrimination in the labor market

A plethora of studies has investigated discrimination of men and women in the labor market. Riach and Rich (1987) were the first to run a correspondence study on gender discrimination and found evidence of discrimination in a subset of the occupations considered. Two recent studies were conducted in France, finding evidence that employers discriminate against younger women for high-skilled jobs (Petit, 2007), and in Belgium where women are discriminated against when applying for jobs which imply a promotion in comparison with their current position (Baert et al., 2016). Azmat and Petrongolo (2014) survey the experimental literature studying gender discrimination. They also consider gender differences in individual preferences and examine whether interactions in groups can explain differences in labor market outcomes. Baert (2017) catalogues correspondence studies on discrimination conducted after 2004 and concludes that the evidence of gender discrimination is mixed and depends on the occupations considered. Researchers have also tried to identify the sources of gender discrimination with the help of correspondence tests, audit studies, and vignette designs. Many studies focus on two factors, namely the *gender ratio* in a profession as well as the *status* of the profession. Table 1 provides an overview of these studies.

The prevalence of one gender in a profession can cause stereotyping, resulting in a more favorable outcome for the dominant gender in the occupation (Booth and Leigh, 2010). Psychological evidence suggests that individuals working in a job where one gender is prevalent think that success in this job requires characteristics typical of that gender (Schein, 1973). As can be taken from the upper part of Table 1, there is evidence of discrimination against men in female-dominated occupations and against women in male-dominated occupations (Levinson, 1975; Riach and Rich, 2006), as well as against men in occupations with a balanced ratio of men and women (Riach and Rich, 2006). However, some studies report no evidence of discrimination against women in some of the male-dominated occupations (Carlsson, 2011; Riach and Rich, 1987; Weichselbaumer, 2004), no evidence of discrimination against men at least in some female-dominated professions (Booth and Leigh, 2010; Carlsson, 2011; Riach and Rich, 1987; Weichselbaumer, 2004), and no evidence of discrimination in some of the gender-neutral occupations (Carlsson, 2011). Similarly, the evidence from early vignette studies on the influence of an occupation's gender-type on discrimination is mixed (e.g., Cash et al., 1977; Sharp and Post, 1980).

We take from this overview that the relationship between gender composition and discrimination deserves closer attention. Note that our sample allows us to compare occupations with a share of female employees between almost 0 and 100 percent.

Table 1: Field experiments and vignette studies on gender discrimination: The effects of the gender ratio and status

Study and Method	Location and Date	Occupations	Finding
Gender ratio:			
Booth and Leigh (2010) (CS)	Brisbane, Melbourne, Sydney, Australia (2007)	Data-entrist (f, 85%), waiters (f, 80%), customer service employees (f, 68%), salesmen (f, 69%)	Discr. against men for data-entry and waitstaff, no discr. for customer service and sales
Levinson (1975) (AS, telephone)	Atlanta, USA (1974)	Mostly secretaries/receptionists (f), mostly security guards/officers/managers/skilled workers (m)	Discr. against men in (f) occupations and against women in (m) occupations, discr. against men more pronounced
Cash et al. (1977) (V, personnel managers)	East Coast, USA (1974)	Car salespersons and hardware clerks (m), motel desk clerks and photography assistants (n), telephone operators and receptionists (f)	Discr. against men in (f) occupations and against women in (m) occupations, no discr. in (n) occupations
Glick et al. (1988) (V, business managers)	Wisconsin, USA (1986)	Sales managers for heavy machinery (m), administrative assistants in bank (n), dental receptionists/secretaries (f)	Discr. against men in (f) occupations and against women in (m) occupations, no discr. in (n) occupations, gender-related individuating information mediates discr.
Riach and Rich (2006) (CS)	England (2003)	Chartered accountants (n, 31%), computer analysts/ programmers (n, 21%), engineers (m, 5%), secretaries (f, 97%)	Discr. against women in (m) occupations and against men in (f) and (n) occupations, discr. against men more pronounced
Riach and Rich (1987) (CS)	Victoria, Australia (1983-1986)	Management accountants (m, 9%), computer programmers (m, 23%), analyst programmers (m, 23%), computer operators (NI), industrial relations officers (NI), clerical workers (f, 68%), gardeners (m, 13%)	Discr. against women for computer programmers and gardeners, for others no discr.
Sharp and Post (1980) (V, personnel managers)	Midwest, USA (1980)	Sports reporters (m, 90%), fashion reporters (f, 98%)	No discr. for sex-incongruent professions
Weichselbaumer (2004) (CS)	Vienna, Austria (1998-1999)	Network technicians (m, 13%), computer programmers (m, 13%), accountants (f, 77%), secretaries (f, 97%)	Discr. against masculine and feminine women for network technicians and against men for secretaries, for others no discr. between men, feminine women, and masculine women

Table 1 (Continued): Field experiments and vignette studies on gender discrimination: The effects of the gender ratio and status

Study and Method	Location and Date	Occupations	Finding
Status:			
Firth (1982) (CS)	England (1978)	Articled clerks and qualified accountants working for professional accounting firms (noncareer jobs), unqualified personnel, qualified accountants working in industry and financial jobs (career jobs)	Discr. against women in qualified accountant positions in industry and in financial jobs; no discr. for all other occupations
Neumark et al. (1996) (AS, in person)	Philadelphia, USA (1994)	Waiters (High, medium and low price/earnings)	Discr. against women in expensive restaurants (job offers and interview invitations) and against men in inexpensive restaurants (only w.r.t. job offers); no discr. in medium-price restaurants
Gender ratio & status:			
Albert et al. (2011) (CS)	Madrid, Spain (2005-2006)	Salesmen (m, l: 21%, h: 30%), accountants (n, l: 46%, h: 49%, secretaries (f, l: 70%, h: 67%)	h: discr. against men in (f) occupation, no discr. in (m) and (n) occupations, l: discr. against men in (f) and (n) occupations, no discr. in (m) occupation
Carlsson (2011) (CS)	Stockholm, Gothenburg, Sweden (2005-2006)	h: Preschool (f), lower (n) and upper secondary teachers in maths/science (n) and languages (n), accountants (f), nurses (f), computer professionals (m), mi: construction workers (m), motor-vehicle drivers (m), business sales assistants (n), shop sales assistants (f), l: restaurant workers (f), cleaners (f)	Discr. against men for restaurant workers, accountants, business sales assistants, preschool teachers and against women for construction workers; no discr. for all other occupations
Muchinsky and Harris (1977) (V, students)	Midwest, USA (1976)	high (h), middle (mi) and low (l) scholastic achievement for childcare (f), journalism (n), mechanical engineering accountants (f), nurses (f), computer professionals (m),	Discr. against men for childcare manager (except for h); No discr. for journalism for l and h, but against women for mi; no discr. against women for mechanical engineering
Zhou et al. (2013) (CS)	Beijing, Shanghai, Guangzhou, Wuhan, Shenzhen, Chengdu, China (2010-2011)	Secretaries (f, 72%), software engineers (m, 31%), accountants (n, 38%), marketing professionals (n, 41%), for each low-ranked and high-ranked position	High-ranked positions: discr. against men for (m),(f) and (n) occupations; low-ranked positions: discr. against men in (f) occupation and in marketing, against women in accounting, and no discr. for (m) occupation

Note: The table shows publications that compare discrimination across occupations with different shares of female employees as well as publications focusing on occupations with different social status. CS refers to a correspondence study, AS indicates an audit study while V stands for a vignette design. (f) means that the authors classify the occupation as female dominated, (m) means that they classify an occupation as male dominated, and (n) means they classify an occupation as gender neutral, where the percentage refers to the ratio of female employees. NI abbreviates "not indicated". l, mi, and h represent "low qualification", "middle qualification" and "high qualification," respectively.

Moreover, it has been argued that discrimination is stronger in occupations of higher status and in jobs that are more senior (Azmat and Petrongolo, 2014; Firth, 1982; Neumark et al., 1996; Riach and Rich, 1987; Riach and Rich, 2002; see the middle part of Table 1). However, this claim is in part based on the findings for occupations with different shares of female employees. Thus, there are possible confounds. Some researchers have also looked at the influence of the gender ratio for occupations of varying occupational status (lower part of Table 1). Albert et al. (2011) for Spain and Zhou et al. (2013) for China find more evidence of discrimination against men than discrimination against women across professions of different status and gender ratios. Carlsson (2011) studies call-back rates in 13 occupations in Sweden and estimates the correlation between discrimination and the share of women in an occupation, the skill level required for an occupation, and firm size, which all turn out to be insignificant.

There may be cultural differences with respect to gender roles and stereotypes that explain the divergence of some of the findings with respect to status. But the results may also differ for other reasons. The studies are based on relatively small sets of occupations. In addition, they either rely on a subjective assessment of occupational status or they only consider one proxy for status, such as educational requirements. Given these drawbacks and the small number of studies, the relationship between discrimination and the status of an occupation calls for further scrutiny.

Several other moderators of gender discrimination have been proposed. Larger firms have been found to discriminate less between men and women (Akar et al., 2014). A possible channel is that a more formalized and professional recruitment procedure might counteract gender discrimination. Another possible explanation is that evaluating larger groups of applicants leads to decisions that are less affected by group stereotypes (Bohnet et al., 2016). In addition, discrimination might vary with the labor market conditions. Becker (1957) shows that discrimination can be hard to sustain with competition from non-discriminating firms. In line with this, there is evidence that ethnic discrimination against a certain group of school-leavers is only observed when labor markets are not tight and vacancies are easy to fill (Baert et al., 2015).[5] With our dataset, we can analyze how the size of the firm and the labor market situation moderate gender discrimination.

Identifying discrimination

Traditionally, labor economists have applied regression-based methods to observational data in order to measure discrimination in the labor market (see Altonji and Blank, 1999, for an overview). One important problem of these methods is the possibility that discrimination is overestimated. In these studies, discrimination is the residual after controlling for observable differences in productivity where these productivity differences might be insufficiently measured. For this reason, many researchers have conducted field experiments to study discrimination in the labor market (for re-

[5]Berson (2012) tests the relationship between the intensity of competition between the firms and hiring discrimination in France. For cashiers, the study reports discrimination against men when competition is strong, but no discrimination when competition is weak.

cent overviews see Baert, 2017; Bertrand and Duflo, 2016; Neumark, 2016). The field experiments compare call-back rates, invitations to job interviews, or job offers for fictitious applicants that are equal with respect to most of their characteristics but differ with respect to gender. Alternative approaches based on observational data were also developed, see e.g., Bayard et al. (2003). Economic lab experiments to measure discrimination are summarized by Lane (2016). Some earlier studies by psychologists in which subjects are asked to rate the employment suitability of candidates are surveyed by Olian et al. (1988).

The vignette design allows us to observe the causal effect on the evaluation of being female or male because we randomize the characteristics of the fictitious CVs, just as in field experiments. However, Heckman (1998) and Heckman and Siegelman (1993) have questioned the validity of field experiments that use correspondence methods. Two points of their criticism are particularly important, and we argue that the factorial survey method can address one of them.

First, even if a researcher conducting a correspondence study is successful in making the applicants of the two groups under consideration equal with respect to the observed productivity characteristics, that is, the characteristics mentioned in the written application, a necessary assumption to identify taste-based discrimination is the equality of the average unobserved productivity-related factors of the two groups. If the assumption is not met, differences in outcomes may be due to differences in these unobservable productivity characteristics or due to group membership. Thus, like field experiments the factorial survey method can only identify discrimination that may be taste-based, statistical, or both, but it cannot differentiate between them.

The second point raised by Heckman and Seligman concerns the *variance* of the unobserved productivity variables. Differences in the variance of the unobserved productivity between the two groups under consideration can cause biased estimates of discrimination if the employers' hiring decisions are based on a cut-off rule (see Neumark, 2012, for a detailed discussion). To see this, assume that both the average observed and unobserved productivity are equal across the two groups. Further assume that the variance of the unobserved productivity is higher for one group, say for women. If the researcher chooses a low value for the observable productivity-related characteristics in the written application, women are more likely to have a productivity that exceeds the cut-off than men and are more likely to be invited to an interview than men. In the extreme case, men are never invited when the variance of their unobservable characteristics is low. In contrast, if the researcher designs applications with a high quality of observable productivity characteristics, the men receive more interview invitations, even in the absence of discrimination. Thus, the observed discrimination can be an artefact of the study design.

Neumark (2012) proposes using a heteroskedastic probit model to estimate discrimination in correspondence studies. Using this model not only allows for estimating discrimination if the variance of unobserved productivity characteristics varies between groups, but also for estimating the ratio of these variances. The method has been widely applied (e.g., Neumark et al., 2016; Baert, 2015; Carlsson et al., 2014). It rests on the assumption that the coefficients of the observable

applicant characteristics used for the identification do not vary between the two groups under consideration. This assumption is violated, e.g., if the productivity effects of schooling differ between the two groups. One issue with this method is the discretion it introduces as to the choice of observable characteristics.

We argue that our estimate of discrimination between men and women is robust to Heckman and Seligman's unobserved variance critique under certain assumptions without introducing this discretion. In Appendix A.1 we formalize the argument that we briefly summarize here. Our estimate of discrimination is based on a 10-point evaluation scale and not on a cut-off value such as a binary invitation decision. Thus, as long as the 10-point scale fully covers all the evaluations the employers want to make, the employers' evaluations of the applicants are linear in productivity and our estimate of discrimination is robust to differences in the variance of unobserved productivity components between men and women.[6] This estimate of discrimination does not rely on the additional assumptions of Neumark's (2012) approach. Moreover, differences in the variance of unobserved productivity simply appear in the variance of evaluations.

However, even if our 10-point scale restricts the evaluation of some respondents, differences in the variance of unobserved productivity components between men and women are unlikely to cause our estimate to be biased for the following reason. As suggested in Neumark (2012), one way to address the unobserved-variance critique is to vary the level of applicant characteristics relevant to the hiring decision. Our vignette study design simultaneously varies several applicant characteristics such as the duration of unemployment after leaving school and the average school grade. Significant coefficients show that these characteristics have an effect on the respondent's choice. Hence, our applicants differ with respect to their observable productivity level. In addition, employers from firms of eight different industries and 126 occupations evaluate our vignettes.[7] Since the quality of an applicant can be expected to depend on the requirements of the firms and occupations, we thereby consider a large variation in the quality of applicants.

Testing for statistical discrimination à la Aigner and Cain (1977)

We can use our dataset to test one of the classical models of statistical labor market discrimination proposed by Aigner and Cain (1977). In this model, productivity is unobserved and employers have to form expectations about a worker's true productivity. The expected productivity of a worker is given by the weighted average of the mean productivity of the group to which the worker belongs and a measure of her individual productivity. In other words, since the true productivity is unobserved, employers rely partly on the group information. The weight on the individual productivity indicator is defined by the variance of the real productivity and the variance of the

[6]This only holds true as long as we abstract from the discrete nature of the outcome variable. We deal with the violation of this assumption by estimating a heteroskedastic ordered probit model as a robustness check (see Section 4.2).

[7]In contrast, in the usual correspondence study applications are sent out for a small number of occupations (up to 13 in the studies we are aware of).

error when measuring the productivity indicator instead of the true productivity, and can be interpreted as the reliability of the productivity indicator. Statistical discrimination can arise due to differences in the average group productivities, but it can also arise due to differences in the reliability of the productivity indicator resulting in a worse signal strength (with average group productivities being equal). Intuitively, employers put more weight on the individual productivity indicator and less weight on the group mean for individuals belonging to the group for which the signal is more informative, which leads to an unequal treatment of equally productive applicants whenever the individual productivity indicator differs from the mean group productivity.[8]

Two features of our study design enable us to test whether a difference in the signal strength due to the different precisions of the productivity indicator causes statistical discrimination. First, we vary several applicant characteristics and, thus, we expect to observe applicants that are considerably above and below the average productivity. Second, under the additional assumption that the current share of employed female apprentices is the decisive factor that causes the signal strength to vary, observing the current share of employed female apprentices provides us with a proxy for the signal strength. The idea is that firms with a high share of female employees have a better prior knowledge of the productivity distribution of women and, hence, the signal emerging from the productivity indicator is more precise. If this reasoning is valid, a better-than-average woman applying to these firms should be more highly evaluated than a man who is better than average and who has the same observable productivity characteristics. That is because employers put more weight on the woman's individual signal than on her group signal compared to a man. In contrast, a lower-than-average woman should be evaluated worse than a lower-than-average man who has the same observable productivity characteristics. Finally, the reverse reasoning applies for firms with a low share of female apprentices. We test whether the evaluations of men and women are consistent with this form of statistical discrimination.

3 Study design

The vignette study was embedded in an annual panel survey of firms engaged in apprenticeship training. In cooperation with the Federal Institute of Vocational Education and Training (BIBB), we included the vignettes (as well as questions concerning the firms' recruitment procedures) in the Company Panel on Qualification and Competence Development in the survey wave of 2014 (BIBB Training Panel 14). The BIBB Training Panel provides information about apprenticeships and other training measures of the dual vocational education system and is a unique data source for research on qualifications, firm-based training, and competency development in Germany (Ger-

[8]The assumption of different variances of real productivity between groups is comparable to Heckman and Siegelman's unobserved variance critique. Note, however, that their criticism is based on an observed and an unobserved productivity component, which are both observed by employers, while in Aigner and Cain's model employers have to form an expectation.

hards et al., 2016).[9] Respondents are firm owners and human resource managers.

The sample of the BIBB qualification panel is obtained from the database of firms from the Federal Employment Agency. The statistical population of the study encompasses all establishments, with at least one employee subject to mandatory social insurance contributions in Germany.[10] The BIBB provides us with sampling weights for our set of firms to correct for imperfections of our sample with respect to the full set of firms with apprenticeship positions in Germany. Thus, we can make statements about the population of all firms in Germany. In addition, we ensure that the sample and the target population are identical by surveying the human resource managers who take the actual hiring decisions in firms. This has been shown to matter for the external validity of vignette studies.[11]

Respondents were first given a number of questions regarding the hiring process. Then they were asked to evaluate short descriptions of applicants (vignettes) for an apprenticeship in the occupation for which their organization trains most people. The managers faced vignettes that were structured like CVs. Thus, their task resembled a task that they perform regularly in their job. The respondents evaluated the vignettes in computer-assisted personal interviews without the interviewer (CAPI). In the vignettes, we randomly varied the applicants' attributes (dimensions). This allows for a detailed study of the relative importance of a number of applicant characteristics simultaneously. Moreover, it is possible to measure the influence of a single variable and of combinations of variables on the evaluations. For example, we can analyze how the effect of being a man varies with the quality of the application. All respondents had to evaluate five vignettes, each of them describing a fictitious applicant, with respect to whether he or she would reach the next step of the hiring procedure. Answers had to be provided on a scale from 1 (very unlikely) to 10 (very likely).

In contrast to correspondence tests, the factorial survey method relies on hypothetical, not real decisions. Hence, the answers can differ from actual behavior for many reasons. For instance, discrimination arising for statistical reasons may be underestimated by our survey if some employers refrain from statistical discrimination when their evaluations do not have any payoff consequences. On the other hand, taste-based discrimination may be more pronounced in the survey, because it is not costly to discriminate. Moreover, employers may not think about a fictitious hiring decision as much as about a real decision. Finally, the answers might be affected by demand effects when respondents try to please the interviewer or provide socially acceptable answers.[12]

[9]The panel provides information about the structure and development of the firms' qualification measures as well as firm-specific labor demand and supply. It encompasses the recruitment behavior of firms and their competence requirements. The panel also provides information on the structural characteristics of the firms. More information about the BIBB qualification panel can be found under https://www.bibb.de/en/1482.php.

[10]This excludes marginally employed employees and some other forms of employment, e.g., public officials.

[11]Analyzing applications for citizenship in Switzerland, Hainmueller et al. (2015) investigate the external validity of a vignette design by comparing its findings to actual choices. They observe that the vignette design captures actual decision-making well, as long as the characteristics of the survey sample and the target population are closely matched.

[12]In principle, this can bias not only our baseline findings, but also the findings of our correlational analysis. For

Table 2: Example of a vignette as shown to the respondents (translated from German)

Gender: Female	**Education:** Intermediate school-leaving degree 2012
Date of birth: Nov. 3, 1995	**Final grade point average:** 3.4
Father's occupation: Employee in firm	**Social behavior according to school reports:** Mostly good
Mother's occupation: Elderly care nurse	**Unexcused absent days:** Three
Current activity: Temporary job (since Dec. 2013)	**Activity since finishing school:** One-year prevocational training and apprenticeship (discontinued)

Note: The respondents were asked: *"How likely is it that this applicant is invited to the next step of the recruitment process?"* Possible answers were between 1 (very unlikely) to 10 (very likely).

In order to limit the concern that demand effects invalidate our results, respondents were not asked directly, for example, for the role of gender or grades for their evaluations, but only indirectly by having to evaluate fictitious applicants. Demand effects are also minimized by the fact that the five fictitious applicants that every respondent received from us were of the same gender (Lahey and Beasley, 2018; Phillips, 2016). At the same time, observing five evaluations by each employer allows us to obtain a large sample. In total, 3,450 firms participate in the panel. Out of these firms, we randomly selected 680 firms to take part in our survey. Hence, we attempted to collect 3,400 evaluations of fictitious applicants. This allows us to analyze how gender discrimination varies with important moderator variables.

Table 2 displays an example of a vignette as it was shown to the respondents (translated from German).[13] Applicants have completed or plan to complete the intermediate schooling degree. All other dimensions were varied. The year of birth was either 1993, 1995 or 1997 such that the applicants were aged 16, 18 or 20 at the start of the apprenticeship. For those born in 1993 and 1995, we varied the information about their activities since leaving school. Either they participated in a one-year pre-vocational training, followed by an apprenticeship that was discontinued, or they did not provide any information on their activities since leaving school two or four years ago.[14] For the applicants born in 1993 and 1995, we further indicated that they were currently working in a temporary job. For the applicants born in 1997, we indicated that this was not applicable since they were still in school.

The average grades from school were either 2.8 or 3.4. Both are satisfactory and typical of

instance, the desire to give socially acceptable answers might differ between occupations with different shares of female apprentices.

[13]The vignette in German and the introductory text shown to the respondents can be found in Section 1 of the Supplementary material.

[14]Pre-vocational training measures are government-sponsored programs aimed at preparing students for apprenticeships. The programs are non-selective. Baethge et al. (2007) and Fitzenberger et al. (2015) provide a description of the various measures of the transition system and their economic significance.

school leavers with an intermediate degree, but a grade point average (GPA) of 2.8 is clearly better than 3.4.[15] Applicants get grades for their social behavior in school, and we varied this between very good and medium. The transcript from school also includes absent days without excuse, and we varied this between none and three days.

Finally, the profession of the father was either warehouse clerk, insurance salesman, teacher, or employee in the respective firm. Many firms have rules according to which children of employees automatically pass the first selection step, and we therefore expect these applicants to pass the hurdle with a higher probability. The other three professions differ with respect to salaries and educational requirements.[16] The mother had one of two professions of similar status, salary, and required education, namely nursery-school teacher and elderly-care nurse.[17]

4 Results

Our main variable of interest is the evaluation of the fictitious applicants on a 10-point scale. The response rate of the 680 firms selected for our survey is more than 98 percent with only 47 out of 3,400 evaluations of fictitious applicants missing. We exclude from the analysis respondents who indicate that they do not know the recruitment procedure of their firm, respondents who state that their firm does not have a recruitment procedure, and respondents who report that their firm does not review and examine the application documents. Overall, this leaves us with a sample of 3,164 evaluations made by 636 respondents. Since the primary purpose of our study is to measure the amount of discrimination in the German apprenticeship market, we state all regression results based on the estimations using the sampling weights. The graphical evidence and the corresponding t-tests are based on the unweighted data.[18]

4.1 Is there a gender difference in evaluations?

For a first impression of the evaluations provided by the respondents, Figure 1 shows the distribution of evaluations for male and female applicants. Apart from indicating that the respondents use the full scale of possible evaluations, the figure reveals a difference in the evaluation of men

[15]Students are not allowed to finish more than two courses with a grade of 5, and have to repeat the school year otherwise. Given this constraint, an average of 3.4 is a relatively poor GPA. On the other hand, students with a GPA higher than 2.8 often decide to get a high school degree (Abitur).

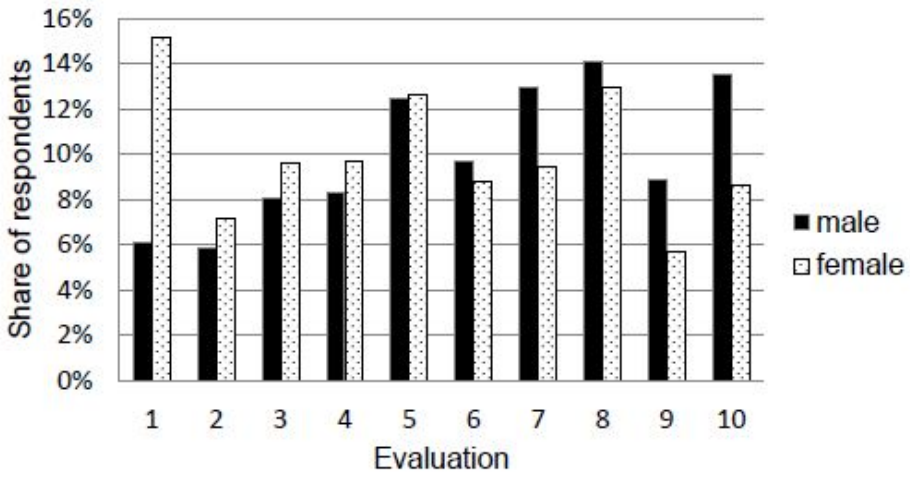
[16]In Germany a warehouse clerk earned, on average, between 12.98 and 15.55 Euro per hour in 2014, while a teacher earned 24.78, and an insurance salesman earned between 23.74 and 28.98 Euro (Federal Statistics Office Destatis, 2016). Working as a teacher requires a university degree in Germany while insurance brokers and (to a lesser degree) warehouse clerks usually complete an apprenticeship.

[17]Nursery-school teachers earned, on average, between 13.60 and 16.43 Euro in Germany in 2014 while elderly-care nurses earned 14.42 Euro (Federal Statistics Office Destatis, 2016).

[18]We conduct all estimations both with and without the sample weights. Most results are qualitatively unchanged. For results without sample weights see the earlier working paper version (Kübler et al., 2017). For a comparison of the weighted and the unweighted dataset see the Supplementary material, Section 2.

and women, with more frequent positive evaluations of men compared to women. Women receive bad evaluations (1–5) more often than men, while men receive good evaluations (6–10) more often than women. The bars for the worst (1) and the best evaluation (10) display the largest differences between women and men. For instance, 241 female applicants receive an evaluation of 1, indicating that an invitation to the next recruitment step is very unlikely, while only 97 male applicants receive this evaluation. Only 137 female applicants receive an evaluation of 10, indicating that the invitation is very likely, while 213 men receive this evaluation. Male applicants receive an average evaluation of 6.14 while female applicants receive an average evaluation of 5.21 such that the male applicants' evaluations are, on average, 0.93 points better than those of female applicants. This difference in means is significant (t-test, $p<0.001$).

Figure 1: Evaluation of fictitious applicants by gender



Note: Fraction of each evaluation for male and female applicants. Evaluations could be made on a scale from 1 to 10. The number of observations is 3,164.

Our vignettes are chosen such that the applicant characteristics are equal for men and women in expectation. Thus, if the randomization worked, the observed mean discrimination against women should be unchanged if we regress the evaluations on a gender dummy and all other vignette dimensions (applicant's age, mother and father's occupation, average grade, social behavior, number of absent days unexcused, and the gap after leaving school).[19] Since one respondent evaluated five vignettes, we use a random-effects (RE) estimation and take into account the dependencies of the answers at the level of the respondents (firms). We also cluster the standard errors at the respondent-level, make use of the sampling weights, and control for the position in which a vignette is shown within the order of the five vignettes.[20] We find that women are evaluated 0.91

[19]In Section 3 of the Supplementary material we report on a comprehensive analysis indicating that the randomization was successful.

[20]In order to both use RE and the sampling weights, we use a multilevel mixed-effects model.

Table 3: Baseline results

	Coefficient	SE
Female	-0.91***	0.174
Age 18	-1.27***	0.253
Age 20	-1.00***	0.307
No info about gap	-0.02	0.171
Worse GPA	-0.65***	0.112
Intermediate social behavior	-0.51***	0.137
Three absent days	-0.76***	0.152
Mother elderly-care nurse	0.25	0.238
Father's occupation		
Insurance salesman	0.07	0.172
Teacher	-0.05	0.207
Employee in firm	0.02	0.287
Constant	8.61***	0.393
Observations	3164	

Notes: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. Coefficients and standard errors are obtained from a RE estimation of the evaluation on the listed variables and dummies for vignette position using the sampling weight. Standard errors are clustered at the respondent-level. Omitted categories are aged 16, information about gap years, warehouse clerk, nursery-school teacher, better GPA, good social behavior, no absent days and first shown vignette.

points lower than men, which is highly statistically significant (see Table 3). It is also economically relevant: Assuming a linear effect of an applicant's GPA on the evaluation, being a man instead of a woman makes up for more than one grade level from school. The only vignette dimensions that have a stronger impact on the respondents' evaluations are a two-year and a four-year gap after leaving school. Since applications from older applicants indicate that they had issues at their career entry, it is not surprising that this variable is relevant for the employers' evaluations.[21] Besides, Table 3 indicates that several other applicant characteristics matter for the employers' evaluations, and the coefficients have the expected signs. This supports the validity of the design.

Do these results regarding discrimination hold up to Heckman and Seligman's unobserved variance critique? A difference in the variance of unobserved productivity characteristics between men and women that results in a difference in the variance of productivity as assessed by the employers should be reflected in a difference in the variance of evaluations. We find that the

[21]Kübler et al. (2018) conducted a field experiment to investigate the effect of a gap after school, informal employment, and training measures completed during the gap on the probability of being invited to an interview for an apprenticeship. They find that a gap of two years is not detrimental for the applicants who have all worked in an informal job.

variance of the evaluations is statistically significantly larger for women than for men.[22] Such a difference can cause the results to be biased unless the study relies on the estimation procedure proposed by Neumark (2012). Our own estimate of discrimination between men and women is robust to the finding of differences in the variance due to the fine-grained evaluations of the applicants that we collected. Note that an unbiased estimate requires that the 10-point scale fully captures the employers' valuations. If respondents wanted to evaluate some women worse and some men better than possible with the scale, our mean estimate would underestimate the true penalty in evaluations for the women. However, it is unlikely that we have set a particularly low or high standard when designing the vignettes, since gender seems to be the main factor leading to extreme valuations: Weaker male applicants are only seldom evaluated poorly (only about 5 percent of men receive an evaluation of 1 or 2), while even good female applicants are not able to score high (only between 5 and 8 percent of women receive an evaluation of 9 or 10, respectively).

4.2 Applicant quality and statistical discrimination

Gender discrimination may depend on the perceived ability of an applicant. We study this possibility in two different ways, and then test the model of statistical discrimination by Aigner and Cain (1977).

First, we compare gender discrimination between applicants with good (2.8) and medium (3.4) grades. The average evaluation is 6.54 and 5.58 for men and women with the better GPA, and 5.71 and 4.83 for men and women with the poorer GPA, respectively. Thus, applicants with the higher GPA are evaluated better than the applicants with the lower GPA. At the same time, there is a gender difference in evaluations for both levels of the GPA, which does not vary much between grade levels. Using a t-test, we find that the gender difference is highly statistically significant for applicants with the better GPA (t-test, $p<0.001$) as well as for applicants with the poorer GPA (t-test, $p<0.001$) and comparable in size (t-test, $p=0.736$).

Second, we can compare the gender difference between applicants grouped according to how they are evaluated by the respondents. Figure 1 suggests that the gender effect is more pronounced for applicants who are evaluated as very likely to be invited to the next step and for those evaluated as very unlikely, and that the gender difference is smaller for intermediate applications. To test this, we define good [bad] applicants as those who are evaluated as good [bad] and use a RE ordered probit estimation to estimate the marginal effect of being female on the probability of observing each of the 10 outcome categories, employing the same controls as before (the vignette dimensions and the vignette position).[23] Being a woman significantly increases the probability of

[22] An F-test of equality of variances strongly rejects the null of equality of variances. We perform Levene's test and the Brown-Forsythe test that are both robust to non-normality of the data, and also find that the variances differ.

[23] See Appendix A.2, Table 5. In all subsequent analyses, we refrain from using RE ordered probit regressions for ease of interpretation.

receiving an evaluation of 1 (the lowest possible evaluation) by 5.3 percentage points. This effect is marginally significant. The gender difference declines over the outcome categories such that the effect of being a female applicant is approximately zero for the intermediate evaluations of 5 and 6. The gender difference is largest for an evaluation of 10 (the highest possible evaluation). Being a woman decreases the probability of this evaluation by 5.0 percentage points. This effect again is marginally significant. Hence, while discrimination is not restricted to only good or only bad applicants, it is highest at the extremes of the distribution of evaluations.

To allow for the fact that the variance of unobserved productivity components differs for men and women and that the 10-point scale might not capture all the evaluations respondents want to make, we estimate a heteroskedastic ordered probit model. We additionally cluster the standard errors at the respondent level. The overall picture remains unchanged.[24]

We can also test whether the observed gender difference in evaluations is due to statistical discrimination caused by differences in the precision of the observable productivity components (Aigner and Cain, 1977). Suppose that a difference in the signal quality of the individual productivity components between men and women is due to the fact that observable productivity components are less precise in measuring real productivity for one gender and is not due to the variance of the real productivity being different.[25] If this difference in the signal quality causes the gender difference in evaluations and if the share of currently employed female apprentices is the decisive factor for differences in signal strength, a better-than-average [worse-than-average] woman applying to firms with a high share of female apprentices should be evaluated better [worse] than a better-than-average [worse-than-average] man with the same observable productivity characteristics applying to the same firms. The opposite holds true for firms with a low share of currently employed female apprentices. Note that our dataset contains self-reported information about the share of women in apprenticeships in a firm, which allows us to perform this test.

We first consider occupations with a share of 67 to 100 percent female apprentices, that is, occupations for which we assume that the signal for the individual component is of higher quality for women than for men. We define a better-than-average applicant as an applicant with the more desirable characteristics of being aged 16, having the better GPA, and zero absent days.[26] For this selected sample, we conduct a t-test on the difference in mean evaluations between men and women. We find the difference to be positive (indicating that men are evaluated better). In turn, we perform a t-test for applicants who we expect to be below average in quality, defined as applicants with less desirable characteristics. Again, we find the difference between men and women

[24]In particular, the effect of being a woman on the probability of receiving an evaluation of 1 is statistically not distinguishable from zero, but it is positive on the probability of receiving an evaluation of 2, 3, and 4 and negative on the probability of receiving an evaluation of 9 and 10.

[25]Note that our results indicate that the variance of real productivity may be higher for women than for men (see Section 4.1). However, this only holds true in the set-up of Heckman (1998) and Heckman and Siegelman (1993) according to which employers evaluate applicants based on their observed and unobserved productivity characteristics.

[26]These three vignette characteristics are the most important according to our baseline results in part 4.1.

to be positive but not significantly different from zero. Repeating this task for male-dominated occupations, we find statistically significant and positive differences in evaluations between men and women not only for better-than-average applicants – as the theory would suggest – but also for worse-than-average applicants – contradicting the theory.

Summing up, we find a statistically significant gender difference in evaluations between men and women for both levels of the GPA. Moreover, discrimination is not restricted to certain parts of the distribution of evaluations. Finally, the gender differences in evaluations are not driven by statistical discrimination based on gender differences in the precision of signals of the individual productivities. Thus, the observed gender discrimination can be due to statistical discrimination caused by differences in the average group productivity, by statistical discrimination if the variance of individual productivity differs between men and women, or by taste-based discrimination.

4.3 Characteristics of occupations and firms

Besides assessing the influence of the vignette characteristics, our dataset allows us to investigate the potential firm- and occupation-related mechanisms behind the observed differential evaluation of men and women. In the remainder of the paper, we will investigate these potential moderators by exploiting the rich dataset. We interact each variable with the gender dummy and control for the dimensions of the vignettes as well as the vignette position in a RE estimation, making use of the sampling weights.[27] In the Supplementary material, tables 5 to 11, we report on the detailed results of all estimations.

4.3.1 Different occupations

We analyze how the observed difference in evaluations between men and women differs between occupation types. Using the survey, we can classify the most common apprenticeship occupation of each firm as "business/administrative," "technical," or "educational/nursing". We regress the evaluation on a gender dummy, dummies for occupations classified as "business/administrative" and "educational/nursing" and the interactions between these two occupation classifications and the gender dummy as described above.

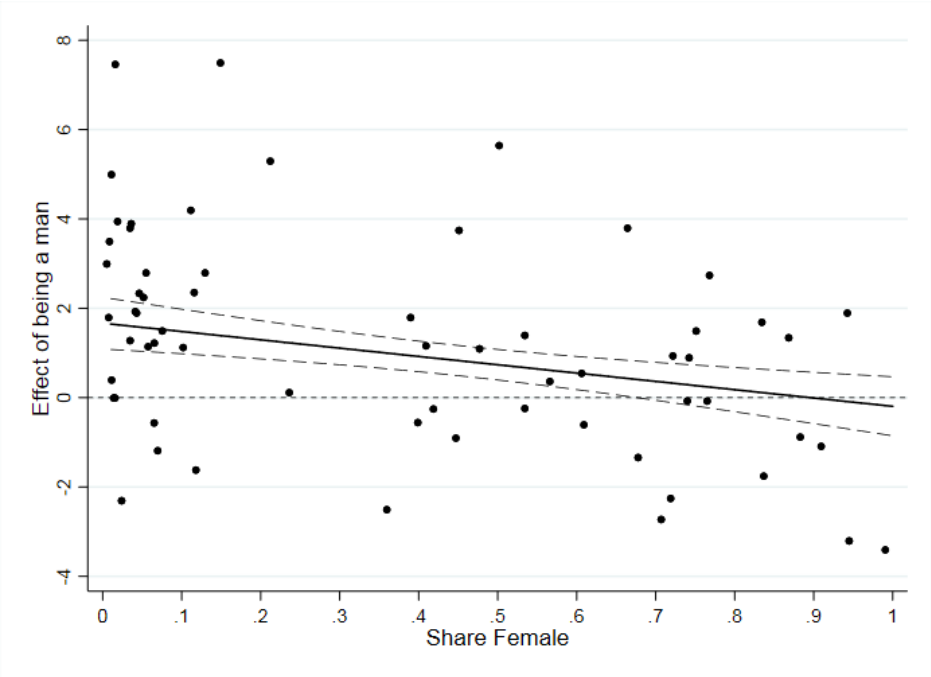
We find that male applicants are evaluated better than female applicants for technical occupations and for educational and nursing occupations. For these occupations being a woman decreases evaluations by 1.23 and 1.07 points, respectively. Although for business or administrative occupations the effect of being a woman is also negative, amounting to 0.44 points, it is not significantly different from zero. There is also variation in discrimination between different industries as reported in Appendix A.3, where we classify firms into eight industries.

[27] As mentioned before, in order to use RE and the sampling weights, we use a multilevel mixed-effects model. Standard errors are clustered at the level of the respondent.

4.3.2 Male- versus female-dominated professions

A potential moderator of gender discrimination is the share of women and men in an occupation. Employers may prefer to hire the dominant gender because they expect this to be conducive to the work atmosphere. To study how gender discrimination depends on the share of employees in the occupation that are female, we add information about the share of women currently in an apprenticeship occupation to our data.[28]

Figure 2: Gender differences in evaluations by share of women



Note: The dots represent gender differences in the evaluation of applicants for apprenticeship occupations with a given share of women. The solid line shows the effect of being a man obtained by a RE regression of the evaluation on a gender dummy, the share of women currently employed in an apprenticeship occupation, the interaction between the share of women and the gender dummy, all other vignette characteristics, and the vignette position using the sampling weight and clustering standard errors at the respondent level. The dashed lines indicate 95% confidence intervals of this effect.

Figure 2 plots the difference in mean evaluations between male and female applicants against the proportion of female employees in the respective occupation in Germany. The dots suggest that the smaller the proportion of female apprentices, the worse the relative evaluation of women compared to men. The figure also shows the results from a RE regression of the evaluation on a gender dummy, the share of women currently employed in an apprenticeship occupation, the interaction

[28]We use the German classification of occupations KldB 2010. The data on the share of female apprentices by profession is provided by the Federal Institute for Vocational Education and Training (BIBB) and the Federal Statistics Office Destatis. See Table 12 in the Supplementary material, for an overview of all external data sources used.

between the share of women and the gender dummy, as well as the other vignette characteristics and the vignette position. There is a negative relationship between the share of current female apprentices and the evaluation of men relative to women. Increasing the share of female apprentices by 50 percentage points is associated with a decrease in the absolute amount of discrimination, namely a decrease in the effect of being a man by 0.93. This effect is highly statistically significant.[29] Moreover, the difference in evaluations between male and female applicants is significantly different from zero at the five percent level for jobs with a share of women below 68 percent, but not significantly different from zero for jobs with a share of women above 68 percent. For example, we find no evidence of discrimination against women for the occupation of nursing where the share of women is 77 percent.[30]

We would like to note that this evidence leaves open whether discrimination is a cause or a consequence. Professions could have become male dominated because of discrimination or because more men have applied to them than women. The finding that there is no discrimination against men in female-dominated professions suggests that men are underrepresented because they do not want to take up these professions. For some professions such as early-childhood education, the absence of discrimination against men could be due to political campaigns aimed at increasing the share of men in such jobs.

4.3.3 High- versus low-status occupations

The literature presents evidence that discrimination against women is more severe for occupations associated with a high social status (for an overview, see Riach and Rich, 2002). We analyze how gender discrimination varies with several variables related to occupational status. In particular, we investigate the relationship between gender discrimination and the average salary, the typical school-leaving qualification required for an apprenticeship, and two indices of occupational status.

Salaries

For the analysis of the salaries, we merge average hourly wages with our professions via the KldB 2010. The wages are national averages provided by the Federal Statistical Office (Destatis). We end up with 117 different hourly wages, ranging from 9.05 Euro per hour to 24.95 Euro per hour.

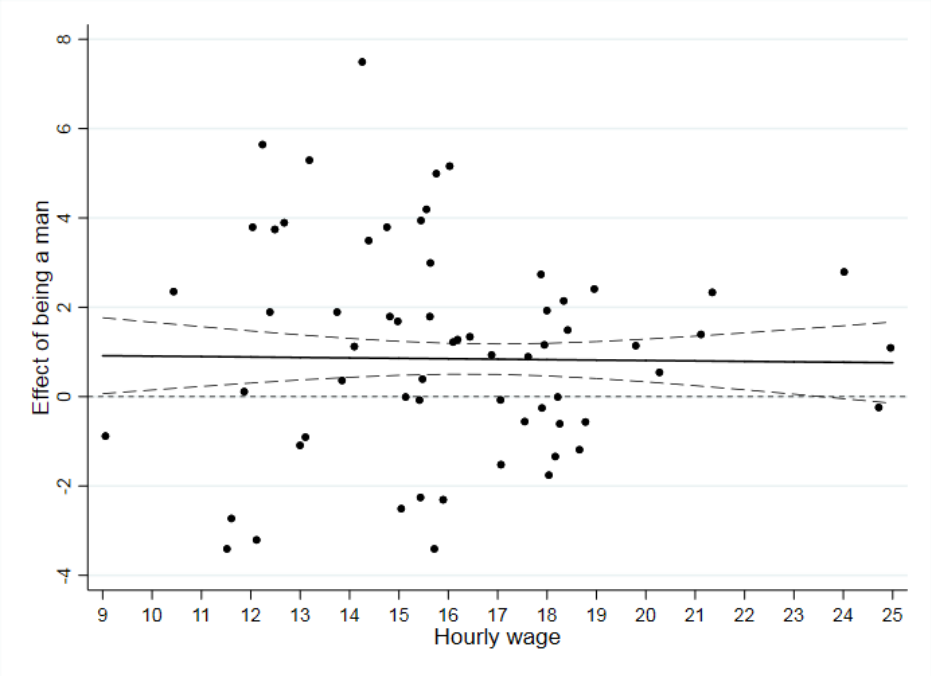
Figure 3 displays how the difference in evaluations between men and women varies with the hourly wage. The graphical evidence suggests no clear linear relationship. The figure also contains the gender effect for different wages obtained by regressing the evaluation on a gender dummy,

[29] If we use the firms' self-reported shares of currently employed female apprentices, an increase of 50 percentage points is associated with a statistically significant increase of 0.77 points in the evaluation of female applicants relative to male applicants.

[30] When we split the occupations according to their share of female apprentices into male-dominated, gender-balanced and female-dominated occupations and interact both the indicator for female-dominated and for male-dominated jobs with the gender dummy in our RE estimation, we neither find significant evidence of discrimination against men nor against women for female-dominated and gender-balanced occupations. In contrast, we find evidence of discrimination against women in male-dominated jobs.

the average hourly wage of an occupation, their interactions, the vignette characteristics, and the vignette position. We see that the negative effect of being a woman slightly decreases in absolute terms with higher wages, and the effect of being a man is not significantly different from zero for very high wages. Overall, there is no clear linear relationship between the amount of discrimination and the salary of a profession.

Figure 3: Gender differences in evaluations by hourly wage of occupation



Note: The dots represent gender differences in the evaluation of applicants for a given wage level. The solid line shows the effect of being a man obtained by a RE regression of the evaluation on a gender dummy, the hourly wage of an occupation, the interaction between the hourly wage of an occupation and the gender dummy, all other vignette characteristics and the vignette position using the sampling weight and clustering standard errors at the respondent-level. The dashed lines indicate 95% confidence intervals for this effect.

To allow for a non-linear effect, we divide the wage distribution at the 0.33-quantile and the 0.66-quantile and interact the gender dummy with indicators for each range of values in a RE regression. There is strong evidence of gender discrimination for firms of all three salary ranges. Hence, we do not find evidence that discrimination systematically varies with the average salary of a profession.

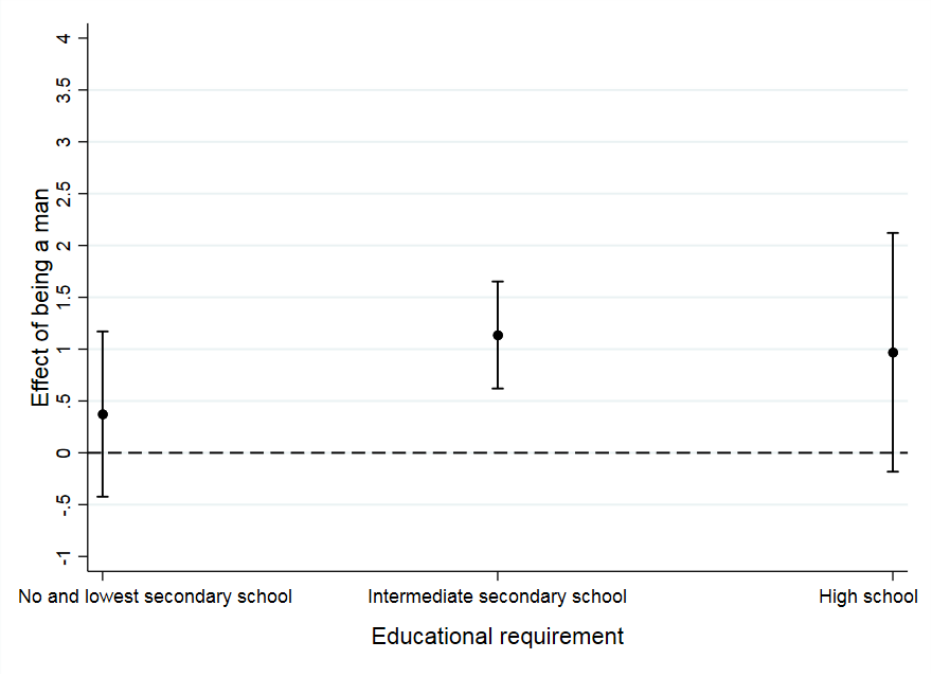
Education requirements

We now consider how discrimination varies with the educational requirements of a profession. After primary school, German students attend different secondary school tracks. These tracks differ with respect to their length and educational orientation. We can differentiate between evaluations made by firms who report that the typical school-leaving qualification of a currently

employed apprentice is (i) no school-leaving qualification or a school-leaving qualification below or equal to lower secondary school (Hauptschule), (ii) intermediate secondary school (Realschule), or (iii) high school (Gymnasium).[31]

Figure 4 shows the results from our correlation analysis. Interacting gender with our indicators of the school-leaving qualification in a regression, we find that the difference in evaluations between men and women is significant at the 5 percent level for the intermediate secondary school, amounting to 1.14 points, but not statistically significant for the other two school-leaving qualifications. When a high school degree is necessary, the difference is also sizable, amounting to 0.97 points but only marginally significant.

Figure 4: Gender differences in evaluations by educational requirement



Note: The figure presents the effect of being a man along with 95% confidence intervals obtained by a RE regression of the evaluation on a gender dummy, the school-leaving qualification required, the interactions between the required school-leaving qualification and the gender dummy, all other vignette characteristics, and the vignette position using the sampling weights and clustering standard errors at the respondent level.

Our dataset can be used to construct another proxy of the typical education required for an apprenticeship, based on a self-reported measure of how many of the firms' apprentices hold each of the various school-leaving qualifications. We build an index based on this distribution and

[31]In the case that a respondent indicates that two school-leaving qualifications are equally likely, we randomly break ties. If a respondent indicates that more than two school-leaving qualifications are equally likely, the observation is discarded. Since only four respondents indicate that their apprentices typically have no school-leaving qualification, we do not distinguish between no school-leaving qualification and the lowest school-leaving qualification ("Hauptschule"). With this procedure, we observe 717 evaluations for "no and lowest secondary school," 1,627 evaluations for "intermediate secondary school" and 579 evaluations for "high school."

report on the analysis in Section 6 of the Supplementary material. Based on this measure, we find that there is discrimination against women for professions and firms with low, medium, and high educational requirements, with no clear differences in the amount of discrimination between these occupations. Thus, overall we find only little evidence that the educational requirements are correlated with the difference in evaluations between men and women.

Occupational Status

Various measures of occupational status have been discussed in the literature (Ganzeboom and Treiman, 2003). We employ the International Socio-Economic Index of occupational status (ISEI-08) which is a socioeconomic measure of occupational status.[32] It is based on the assumption that occupational status determines how education is transformed into earnings.[33] The index was obtained using data from 2002 to 2007 on 200,000 individuals from more than 40 countries (Ganzeboom, 2010). In our sample the ISEI is lowest for unskilled agricultural laborers who have a score of 11.74 and highest for application programmers with a score of 74.66. The average score is 36.92 and applies, for instance, to electrical mechanics.

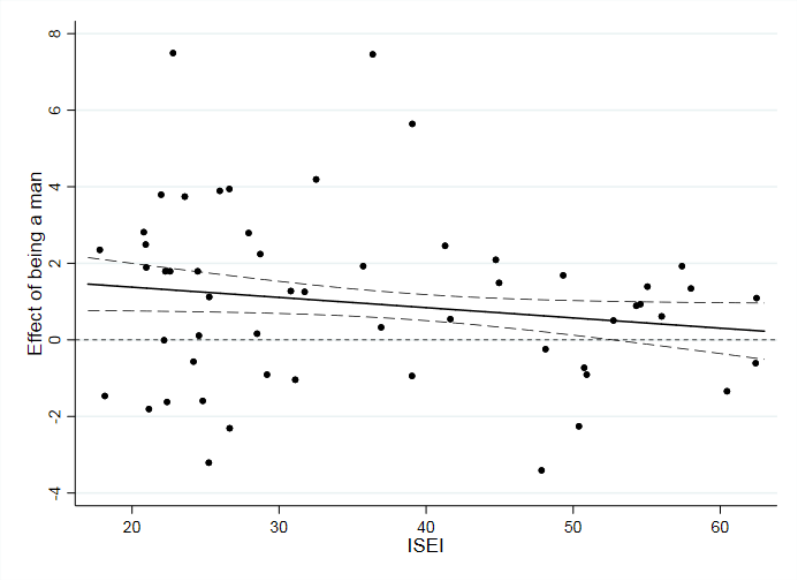
Figure 5 displays the average differences in evaluations between male and female applicants for different values of the index of socioeconomic status. Male applicants are evaluated better than female applicants for the majority of status scores. The dots further suggest a negative relationship between the difference in evaluations of male and female applicants and the socioeconomic status of an occupation. This is in line with the result that we obtain when we interact the socioeconomic measure of occupational status with gender and run a RE estimation on a gender dummy, the ISEI, the interaction between them, the vignette characteristics and the vignette position. There is a sizable negative baseline effect of being female on the evaluations, while an increase in the status of an occupation is associated with a decrease in gender discrimination as indicated by a marginally significant coefficient on the interaction term. In line with this, we find statistically significant discrimination of 1.24 points for waitresses but no evidence of discrimination for software developers (effect size=0.09).

Hence, there is some evidence that status moderates discrimination. In contrast to existing studies (Neumark et al., 1996), we find more severe discrimination against women for occupations

[32]The ISEI-08 is the International Socio-Economic Index of occupational status estimated for the ISCO-08, a classification of occupations based on skill level and skill specialization. We rely on the conversion provided by Ganzeboom and Treiman (2017). Since in our dataset firms are categorized according to the KldB 2010, we first add the ISCO-08 to our data, that is, we assign one unit group of the ISCO-08 (four-digit) to every occupational type of the KldB 2010 (five-digit). We do this by using the conversion key provided by the Federal Employment Agency, see https://statistik.arbeitsagentur.de/Navigation/Statistik/Grundlagen/Klassifikation-der-Berufe/KldB2010/Arbeitshilfen/Umsteigeschluessel/Umsteigeschluessel-Nav.html.

[33]The ISEI of an occupation is found by scaling occupational groups as a mediating variable such that the direct effects of education on earnings are minimized and effects of education on earnings through occupation are maximized. This leads to two coefficients, one linking occupational status to education and one relating occupational status to income that are used as weights to produce status scores based on the average level of education and earnings within an occupation.

Figure 5: Gender differences in evaluations by status of occupation



Note: The dots represent gender differences in the evaluation of applicants for a certain ISEI score. The solid line shows the effect of being a man obtained by a RE regression of the evaluation on a gender dummy, the ISEI, the interaction between the ISEI and the gender dummy, all other vignette characteristics and the vignette position using the sampling weight and clustering standard errors at the respondent-level. The dashed lines indicate 95% confidence intervals for this effect.

that rank lower in the status hierarchy. Note that our analysis includes more than 120 occupations and that we use a measure of status from the literature, in contrast to previous studies that were restricted to small sets of occupations and no validated measures of status. However, it should be kept in mind that we are only considering occupations that do not require a university degree. It could very well be that the picture is reversed for high-status occupations requiring a university degree and for senior management jobs.

4.3.4 Firm size, professionalization of the hiring procedure, and tightness of the labor market

Previous research has found a negative relationship between firm size and discrimination (Akar et al., 2014; Scherr et al., 2013). If the size of the firm is somehow reflected in the individual evaluations made by our respondents, we would expect discrimination to be a decreasing function of firm size. We use the measure of the firm size provided by the BIBB that is based on four categories of one to 19 employees, 20 to 99, 100 to 199, and 200 or more.[34]

Controlling for the vignette characteristics and the vignette position, for applicants evaluated

[34]In the BIBB Training Panel the number of employees indicated by the respondents refers to employees for whom the firm has to make social security payments. Marginally employed persons (defined as those earning less than 450 Euro per month) are not taken into account. Sample sizes per category range from 439 (for 100 to 199 employees) to 1,056 (for 200 or more employees).

by the smallest firms (1 to 19 employees), the disadvantage of female applicants is the biggest and amounts to 1.27 points. The disadvantage is smaller for the other three categories, but it is still sizable and statistically significant, and no linear relationship between firm size and the amount of discrimination emerges.

When we use the number of apprenticeships in the firm instead of the number of employees as a measure of firm size, we again find no clear evidence that discrimination depends on firm size. The coefficient on the interaction term between gender and the number of apprentices is not significantly different from zero. The estimated effect of being female is -0.88 for firms with four apprentices (the median) and -0.79 for the 0.9-quantile (31 apprentices). In both cases, the gender effect is significantly different from zero.

A related variable potentially moderating gender discrimination is the professionalization of the hiring procedure, which is likely to be correlated with firm size. A more professionalized and standardized recruitment process might counteract gender discrimination, for example, because standardized evaluation criteria are applied or because managers with a taste for discrimination have to justify their decisions vis-à-vis others. Of course, the evidence from our vignettes can only pick this up if individual managers who respond to our survey have internalized the features of their firm's recruitment process. We measure a firm's degree of professionalization by the number of steps that a firm's recruitment process typically takes. These include, for example, the job interview, tests based on work-related or school-related questions, intelligence tests or personality tests, or trial work in the firm. Clearly, not all of the steps are required in every firm, but we interpret a lack of many of these steps as a lower level of professionalization. If we interact the number of steps in the recruitment process linearly with gender, we find the coefficient on the interaction term to be -0.09 and not significantly different from zero. There is also no evidence that discrimination depends on the number of steps if we use a dummy for more than the median number of steps (three steps) or a dummy for more than the mean number of steps (four steps).

Finally, we consider whether labor market conditions moderate gender discrimination. Specifically, human resource managers can be expected to discriminate less against women if they face a limited supply of applicants. In this case, the cost of discrimination is likely to be higher in the sense of Becker (1957), because there are not many suitable male candidates. As an indicator of labor market tightness, we look at the percentage of apprenticeships of an occupation in a firm that have remained unfilled in past years (a categorical variable with the categories 0, 25% or less, 25 to 50%, and 50% or more).[35]

We find no evidence of gender discrimination if more than 50 percent of the apprenticeships have remained unfilled (effect size: -0.27), but significantly more and marginally statistically significant discrimination if only 25 to 50 percent of the apprenticeships have remained unfilled (effect size:

[35]The majority of firms have had no unfilled apprenticeship positions in recent years. For this category we observe 2,316 observations. The number of observations for the other categories is lower, e.g., we observe 187 observations for the category 25 to 50%.

-1.42), and statistically significant discrimination at the 5 percent level if less than 25 percent (effect size: -1.48) or no apprenticeships (effect size: -0.84) have remained unfilled in the last three years. Hence, labor market conditions have the expected effect on discrimination.

In summary, we find little evidence that the size of a firm or the degree of professionalization of the recruitment procedure have a systematic influence on discrimination at the early stage of the selection process that we are focusing on. On the other hand, we observe that human resource managers seem to discriminate less when the supply of suitable applicants is very low.

4.4 Disentangling the effects of firms, occupations, and industries

The previous sections were devoted to the analysis of how discrimination varies with several firm-related or occupation-related moderator variables. In this section we attempt to determine which of these variables have explanatory power for the variation in the observed gender difference. Existing field experiments have analyzed the effect of some of these moderators alone, but it is likely that by varying, for example, the share of female employees in an occupation, these studies also vary other moderator variables, such as the status of these occupations, the salary, etc. In fact, the share of currently employed female apprentices and the variables related to status (educational requirements, salaries, status index) are weakly correlated in our sample.[36]

We consider the explanatory power of all our variables of interest in one regression. Thus, we employ interaction terms between the gender dummy and all moderator variables analyzed in the previous subsection. We then regress the applicant's evaluation on our moderator variables and the interaction between these moderators and gender.[37] We additionally control for all other vignette characteristics and the vignette position. For comparability, we standardize all variables such that the coefficients can be interpreted in terms of standard deviations.

In a first step, we include the variables related to status, namely the wage paid, the education required, and the status measure in the RE regression using the weighted data.[38] The results can be found in column (1) of Table 4. It shows that the coefficients on the interaction terms between the hourly wage and gender, the education index and gender, and the ISEI and gender are all not

[36]In particular, the share of female apprentices is weakly positively correlated with the educational requirements of an occupation (r=0.27 with the typical school-leaving degree and r=0.27 using the education index). It is weakly negatively correlated with the hourly wage of an apprenticeship occupation (r=-0.17) and moderately positively correlated with the ISEI (r=0.56).

[37]The moderator variables are the occupation classification of the apprenticeship as technical, business/administrative, or educational/nursing, share of women as apprentices in the firm, average hourly wage of an occupation, educational requirement of an occupation measured by the index, status index (ISEI), firm size measured by the number of apprentices in the firm, professionalization of the human resource management and labor market tightness measured by the fraction of apprenticeships of an occupation in a firm that have remained unfilled. The main results are unchanged if we use the number of employers as the measure of firm size, the classification of firms into industries instead of the classification of apprenticeship occupations into technical, business/administrative, or educational/nursing occupations and education indicators instead of the education index.

[38]For ease of interpretation, we use the education index introduced in Appendix **??** where a higher number stands for more demanding educational requirements.

statistically different from zero. The same holds true for the interaction terms between gender and firm size as measured by the number of apprentices, between gender and the number of steps in the recruitment process, between gender and the dummies for the classes of occupations (business or administrative and educational or nursing), and between gender and the dummies for the fraction of unfilled vacancies. Hence, we find no evidence that any of these variables moderate the gender difference in evaluations.

In contrast, we find a positive coefficient for the interaction between the share of female apprentices and gender. An increase in the share of currently employed female apprentices by one standard deviation goes along with a decrease in the effect of being female on the evaluation by 0.86 points. This decrease is highly significant. It is also relatively large, since it is bigger in absolute terms than the effect of having only an intermediate (3.4) instead of a relatively good GPA (2.8) on evaluations (not reported in the table). Quantitatively, it implies that an increase in the share of female apprentices by 50 percentage points is associated with a 1.34-point increase in female evaluations relative to male evaluations.

Observing no statistically significant correlations of the status-related variables on the gender difference in evaluations might be caused by including different variables for status that are correlated with each other. Indeed, we find that the educational requirement, the average wage in the occupation, and the status index are correlated.[39] Therefore, we re-run our RE estimation including only one of the status-related variables at a time (see columns (2) to (4)). We again find no evidence that any of the status-related variables (salary, educational requirement, ISEI) correlate with the gender differences in evaluations. The coefficient on the interaction term between gender and the share of female apprentices decreases slightly, but it remains positive and sizable. Hence, over all the specifications there is a statistically significant and positive correlation of the share of current female apprentices and the effect of being a female applicant.

We draw a number of conclusions from the analysis presented in this section. The difference in evaluations between male and female applicants is most strongly correlated with the share of female applicants in an occupation. Discrimination against female applicants decreases with the share of female apprentices currently employed. Further analyses based on a sample split for male and female applicants show that most of the difference is due to women being evaluated better when the share of female apprentices increases while the difference is not caused by a worse evaluation of male applicants when the share of women increases.[40] We find no evidence that the variables related to status, such as the average hourly wage of an occupation, an occupation's educational requirements, and the index of occupational status, as well as firm size, the type of occupation, the professionalization of the recruitment process or the labor market tightness can

[39]The wage variable is weakly to moderately correlated with educational requirements (r=0.39 with dummies for education levels, r=0.40 with the education index) and weakly correlated with the measure of occupational status (r=0.39). The correlation between educational requirements and the status index is moderately positive (r=0.42 (r=0.43) using indicators for education (the education index)).

[40]The regressions are reported in Section 7 of the Supplementary material.

explain a significant proportion of the observed gender difference.

Table 4: Disentangling the effects of the moderator variables

	(1)	(2)	(3)	(4)
Fem	-0.84***	-0.76***	-0.84***	-0.83***
	(0.283)	(0.279)	(0.283)	(0.279)
Fem*Share Fem	0.86***	0.74***	0.70***	0.66***
	(0.275)	(0.235)	(0.226)	(0.238)
Fem*Hourly Wage	0.15	0.14		
	(0.232)	(0.179)		
Fem*Education	-0.03		-0.02	
	(0.220)		(0.189)	
Fem*ISEI	-0.12			0.00
	(0.273)			(0.213)
Fem*Firm Size	-0.23	-0.09	-0.20	-0.10
	(0.187)	(0.140)	(0.188)	(0.130)
Fem*Steps	-0.06	-0.05	-0.07	-0.06
	(0.172)	(0.165)	(0.169)	(0.164)
Occupation type				
Fem*Business/administrative	0.06	-0.06	0.14	0.07
	(0.467)	(0.461)	(0.463)	(0.450)
Fem*Educational/nursing	-0.50	-0.62	-0.27	-0.51
	(0.716)	(0.668)	(0.665)	(0.709)
Tightness				
Female*vacancies <=25%	-0.44	-0.50	-0.47	-0.52
	(0.486)	(0.501)	(0.497)	(0.496)
Female* vacancies >25% to 50%	-0.24	-0.23	-0.36	-0.300
	(0.846)	(0.808)	(0.848)	(0.799)
Female* vacancies >50%	-0.40	0.50	-0.54	0.39
	(1.007)	(0.838)	(1.036)	(0.831)
Controls	Yes	Yes	Yes	Yes
Observations	2849	3074	2863	3061

Notes: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. The standard errors in parentheses are clustered at the respondent-level. "Fem" indicates a dummy for female applicants. "Share Fem" indicates the share of currently employed women in an apprenticeship occupation. Education is measured by the education index, and firm size is measured by the number of current apprenticeships in the firm. Controls are the vignette characteristics (unemployment spell, information on unemployment spell, father's occupation, mother's occupation, GPA, social behavior, number of absent days) and the vignette position. The omitted categories are the class of technical apprenticeship occupations, no unfilled vacancies, aged 16, information about gap years, warehouse clerk, elderly-care nurse, better GPA, good social behavior, no absent days and first shown vignette. The table reports only the coefficients on the interaction terms.

5 Conclusion

We document that female job applicants are evaluated worse than male applicants by human resource managers from a large sample of German firms. No matter how we split our sample by observable characteristics of firms, occupations, or job applicants, discrimination against women is almost always observed, albeit to varying degrees. The observed discrimination can help explain existing gender gaps in employment rates at career entry and in particular in apprenticeship positions. Our findings cannot be due to differences in the expected variance of the unobserved productivity of men and women. The observed gender differences are also unlikely to represent statistical discrimination caused by the productivity signals of the underrepresented gender being less reliable.

Both technical professions and professions in education and nursing display a significant penalty for female applicants, in contrast to business and administrative professions. Also, women fare significantly worse than men in male-dominated professions. There is no clear relationship between educational requirements and discrimination. There is also no clear relationship between the salaries of professions and discrimination, but we find that the evaluations of women relative to men improve with the increasing social status of an occupation (as measured by the index ISEI). In tight labor markets, the penalty for women is also low. Finally, there is no correlation between the size of the firm and the structure of the recruitment procedure with the amount of gender discrimination.

When we consider the relationship between gender discrimination and all firm- and occupation-specific moderator variables simultaneously, only the gender ratio in a profession shows a significant correlation with gender discrimination. In particular, women suffer large penalties in male-dominated professions. While we find evidence that the occupational status of an occupation and the labor market situation moderate gender discrimination, this evidence is weaker because it is not robust to the inclusion of all variables. However, the effect of the gender ratio is robust to the inclusion of a large set of moderator variables. This finding regarding the correlation between discrimination and the gender ratio of a profession is consistent with most of the evidence from correspondence, audit, and vignette studies summarized in Table 1.

The factorial survey design does not allow for observing actual hiring decisions and, hence, can only complement findings based on observational and experimental data. Yet, the responses that we have collected from a nationally-representative survey lead to clear differences between the evaluations of our candidates, with variables such as grades and age having the expected effects. We interpret this as pointing toward the validity of our design. One reason for this might be that our survey was conducted as part of a regular panel with a response rate of almost 100 percent, and that we presented the vignettes to individuals for whom evaluating CVs is a part of their everyday tasks.

One unobserved variable that could determine the productivity in certain professions is physical

strength. If employers expect men to be, on average, stronger than women, this could give them an advantage, for example, in the construction industry or in manufacturing where we find the largest differences in evaluations. We would like to make two comments regarding this concern. First, many of the most common professions with a strong male dominance do not require physical strength, such as a mechatronics engineer, and many similar technical occupations. Second, the amount of strength needed for certain tasks is a function of the tools developed for them. "Handle size and tool weight are designed to accommodate the size and strength of men," as stated in a report to OSHA (Occupational Safety and Health Administration) of the Advisory Committee on Construction Safety and Health of the US Department of Labor (1999), which recommends changes necessary to allow women to enter the construction industry. While we cannot exclude that differences in physical strength can explain some of the differences in the evaluations of men and women, at least part of the productivity differential could be removed via appropriate adjustments of the workplace.

Our results indicate that in male-dominated professions, employers are less likely to hire women compared to men. Thus, discrimination is likely to push women into female-dominated jobs. This can perpetuate gender imbalances and contribute to a shortage of labor supply, for example, in technical professions, especially if women anticipate the firms' responses and therefore apply at a lower rate to professions in which they are underrepresented. Given the design of our study, we can only make causal statements regarding the effect of gender on the evaluation of an applicant and not regarding the effect of the gender ratio on discrimination. However, our results are consistent with policies that aim at increasing the number of women in jobs where they are underrepresented. Such policies could change the gender stereotypes attached to these jobs and the atmosphere at the workplace that are often seen as culprits of self-perpetuating gender differentials.

References

[1] Advisory Committee on Construction Safety and Health (1999). Women in the Construction Workplace: Providing Equitable Safety and Health Protection. Submitted to the Occupational Safety and Health Administration (OSHA), Department of Labor, USA. Retrieved from https:// www.osha.gov/doc/accsh/haswicformal.html#ergonomics.

[2] Aigner, D.J., and G.G. Cain (1977). Statistical Theories of Discrimination in Labor Markets. ILR Review 30.2 : 175–187.

[3] Akar, G., B. Balkan, and S. Tumen (2014). Overview of Firm-Size and Gender Pay Gaps in Turkey: The Role of Informal Employment.

[4] Albert A., L. Escot and J.A. Fernández-Cornejo (2011). A Field Experiment to Study Sex and Age Discrimination in the Madrid Labour Market. International Journal of Human Resource Management 22, 351–375.

[5] Altonji, J.G. and R.M. Blank (1999). Race and Gender in the Labor Market. In: Ashenfelter, O. and D. Card (eds.). Handbook of Labor Economics. Vol. 3, chapter 48.

[6] Ambuehl, S., and A. Ockenfels (2017). The Ethics of Incentivizing the Uninformed: A Vignette Study. American Economic Review, 107(5), 91–95.

[7] Ambuehl, S., M. Niederle and A. E. Roth (2015). More Money, More Problems? Can High Pay Be Coercive and Repugnant? American Economic Review, 105(5), 357–60.

[8] Azmat, G. and B. Petrongolo (2014). Gender and the Labor Market: What Have We Learned From Field and Lab Experiments? Labour Economics 30, 32–40.

[9] Baethge, M., H. Solga and M. Wieck (2007). Berufsbildung im Umbruch - Signale eines überfälligen Aufbruchs. Netzwerk Bildung. Berlin: Friedrich-Ebert-Stiftung.

[10] Baert, S. (2015). Field Experimental Evidence on Gender Discrimination in Hiring: Biased as Heckman and Siegelman Predicted? Economics: The Open-Access, Open-Assessment E-Journal 9, 25.

[11] Baert, S. (2017). Hiring Discrimination: An Overview of (Almost) All Correspondence Experiments Since 2005. GLO Discussion Paper, 61.

[12] Baert, S. and A.S. De Pauw (2014). Is Ethnic Discrimination Due to Distaste or Statistics? Economics Letters, 125 (2), 270–273.

[13] Baert S., A.S. De Pauw and N. Deschacht (2016). Do Employer Preferences Contribute to Sticky Floors? ILR Review 69, 714–736.

[14] Baert S., B. Cockx, N. Gheylem C. Vandamme (2015). Is There Less Discrimination in Occupations Where Recruitment Is Difficult? ILR Review 68, 467–500.

[15] Bayard, K., J. Hellerstein, D. Neumark and K. Troske (2003). New Evidence on Sex Segregation and Sex Differences in Wages from Matched Employee-employer Data. Journal of Labor Economics, 21(4), 887–922.

[16] Bayard, K., J. Hellerstein, D. Neumark and K. Troske (2003). New Evidence on Sex Segregation and Sex Differences in Wages from Matched Employee-employer Data. Journal of Labor Economics, 21(4), 887–922.

[17] Becker, G.S. (1957). The economics of discrimination. University of Chicago Press, Chicago.

[18] Behr, A. and K. Theune (2016). The Gender Pay Gap at Labour Market Entrance: Evidence for Germany. International Labour Review. Accepted Author Manuscript, doi:10.1111/ilr.12037.

[19] Berson B. (2012). Does Competition Induce Hiring Equity? Documents de travail du Centre d'Economie de la Sorbonne, 12019.

[20] Bertrand, M. and E. Duflo (2016). Field Experiments on Discrimination. Handbook of Economic Field Experiments, 1, 309–393.

[21] BIBB [Bundesinstitut für Berufsbildung] (2017). Datenreport zum Berufsbildungsbericht 2017, Table A7.3-4. Bonn: BIBB. Online: https://www.bibb.de/dokumente/pdf/bibb_datenreport_2017.pdf, last access: 11/03/2017.

[22] Blau, F.D. and L.M. Kahn (2007). The Gender Pay Gap: Have Women Gone as Far as They Can? The Academy of Management Perspectives, 21(1), 7–23.

[23] Blommaert, L., M. Coenders, F. van Tubergen (2014). Ethnic Discrimination in Recruitment and Decision Makers' Features: Evidence from Laboratory Experiment and Survey Data using a Student Sample. Social Indicators Research 116(3),731–54.

[24] Bohnet, I., A. van Geen, and M. Bazerman (2016). When Performance Trumps Gender Bias: Joint Versus Separate Evaluation. Management Science 62(5), 1225–1234.

[25] Booth, A. and A. Leigh (2010). Do Employers Discriminate by Gender? A Field Experiment in Female-Dominated Occupations. Economic Letters 107, 236–238.

[26] Bureau of Labor Statistics, U.S. Department of Labor, Occupational Employment Statistics, www.bls.gov/oes/, last access: 11/06/2017.

[27] Carlsson, M. (2011). Does Hiring Discrimination Cause Gender Segregation in the Swedish Labor Market? Feminist Economics 17, 71–102.

[28] Carlsson, M., L. Fumarco, and D.O. Rooth (2014). Does the Design of Correspondence Studies Influence the Measurement of Discrimination? IZA Journal of Migration 2014(3), 11.

[29] Cash, T. F., B. Gillen and D.S. Burns (1977). Sexism and "Beautyism" in Personnel Consultant Decision Making. Journal of Applied Psychology, 62, 301–310.

[30] Cohen, S. L., and K.A. Bunker (1975). Subtle Effects of Sex Role Stereotypes on Recruiters' Hiring Decisions. Journal of Applied Psychology, 60, 566–572.

[31] Correll, S. J., S. Benard and I. Paik (2007). Getting a Job: Is there a Motherhood Penalty? American Journal of Sociology, 112(5), 1297–1338.

[32] Federal Employment Agency (2017). Statistik/Arbeitsmarktberichterstattung: Der Arbeits- und Ausbildungsmarkt in Deutschland – Monatsbericht, Oktober 2017, Nürnberg.

[33] Fershtman, C. and U. Gneezy (2001). Discrimination in a Segmented Society: An Experimental Approach, Quarterly Journal of Economics, 116, 351–377.

[34] Finseraas, H., Å.A. Johnsen, A. Kotsadam and G. Torsvik (2016). Exposure to Female Colleagues Breaks the Glass Ceiling – Evidence from a Combined Vignette and Field Experiment, European Economic Review, 90, 363–374.

[35] Firth, M. (1982). Sex Discrimination in Job Opportunities for Women. Sex Roles, 8 (8), 891–901.

[36] Fitzenberger, B., S. Lickleder er, and M. Zimmermann (2015). Übergänge von der allgemeinbildenden Schule in berufliche Ausbildung und Arbeitsmarkt: Die ökonomische Perspektive. Forthcoming in: Seifried, J., Seeber, S. und Ziegler, B. (Ed.): Jahrbuch der berufs- und wirtschaftspä dagogischen Forschung 2015. Schriftenreihe der Sektion Berufs-und Wirtschaftspädagogik. Opladen: Barbara Budrich.

[37] Ganzeboom, H.B.G. (2010). A New International Socio-Economic Index [ISEI] of Occupational Status for the International Standard Classification of Occupation 2008 [ISCO-08] Constructed with Data from the ISSP 2002-2007; with an Analysis of Quality of Occupational Measurement in ISSP. Paper presented at Annual Conference of International Social Survey Programme, Lisbon, May 1 2010.

[38] Ganzeboom, H.B.G. and D.J. Treiman (2017). International Stratification and Mobility File: Conversion Tools. Amsterdam: Department of Social Research Methodology, http://www.harryganzeboom.nl/ismf/index.htm. Last revision on April 7, 2017.

[39] Ganzeboom, H.B.G. and D.J. Treiman (2003). Three Internationally Standardised Measures for Comparative Research on Occupational Status. Pages 159-193. In: Jürgen H.P. Hoffmeyer-Zlotnik & Christof Wolf (Eds.), Advances in Cross-National Comparison. A European Working Book for Demographic and Socio-Economic Variables. New York: Kluwer Academic Press.

[40] Gerhards, C., S. Mohr and K. Troltsch (2016). BIBB Training Panel - An Establishment Panel on Training and Competence Development 2014. gwa_1.0; Research Data Center at BIBB; Bonn: Federal Institute for Vocational Education and Training. doi:10.7803/371.14.1.2.10.

[41] Glick, P., Zion, C., and C. Nelson (1988). What Mediates Sex Discrimination in Hiring Decisions? Journal of Personality and Social Psychology 55 (2), 178–186.

[42] Hainmueller, J., D. Hangartner, and T. Yamamoto (2015). Validating vignette and conjoint survey experiments against real-world behavior. Proceedings of the National Academy of Sciences 112.8, 2395–2400.

[43] Heckman, J.J (1998). Detecting Discrimination. The Journal of Economic Perspectives 12.2: 101–116.

[44] Heckman, J.J. and P. Siegelman (1993). The Urban Institute Audit Studies: Their Methods and Findings.

[45] Heilman, M. E. (1984). Information as a Deterrent against Sex Discrimination: The Effects of Applicant Sex and Information Type on Preliminary Employment Decisions. Organizational Behavior and Human Performance 33(2), 174–86.

[46] Kleven, H. J., C. Landais, and J.E. Sogaard (2017). Children and Gender Inequality: Evidence from Denmark. Unpublished manuscript, London School Econ.

[47] Kübler, D., J. Schmid, and R. Stüber (2018). Take Your Time to Grow: A Field Experiment on the Hiring of Youths in Germany. Mimeo.

[48] Kübler, D., J. Schmid, and R. Stüber (2017). Be a Man or Become a Nurse: Comparing Gender Discrimination by Employers Across a Wide Variety of Professions. WZB Discussion Paper SP II 2017-201.

[49] Kunze, A. (2003). Gender Differences in Entry Wages and Early Career Wages. Annales d'Èconomie et de Statistique, 245–265.

[50] Kunze, A. (2005). The Evolution of the Gender Wage Gap. Labor Economics 12(1), 73–97.

[51] Lahey, J. N. and R. Beasley (2018). Technical Aspects of Correspondence Studies. In: Gaddis S. (eds) Audit Studies: Behind the Scenes with Theory, Method, and Nuance. Methodos Series (Methodological Prospects in the Social Sciences), vol 14. Springer, Cham.

[52] Lahey, J. N. and D. R. Oxley (2017). Discrimination at the Intersection of Age, Race, and Gender: Evidence from a Lab-in-the-Field Experiment.

[53] Lane, T. (2016). Discrimination in the Laboratory: A Meta-analysis of Economics Experiments. European Economic Review 90, 375–402.

[54] Levinson, R.M. (1975). Sex Discrimination And Employment Practices: An Experiment With Unconventional Job Inquiries. Social Problems 22, 533–543.

[55] List, J. (2004). The Nature and Extent of Discrimination in the Market Place: Evidence from the Field. Quarterly Journal of Economics 119, 49–90.

[56] Marini, M., and P. Fan (1997). The Gender Gap in Earnings at Career Entry. American Sociological Review, 62(4), 588–604.

[57] Muchinsky, P. M. and Harris, S. L. (1977). The Effect of Applicant Sex and Scholastic Standing on the Evaluation of Job Applicant Resumes in Sex-typed Occupations. Journal of Vocational Behavior, 11(1), 95–108.

[58] Neumark, D. (2012). Detecting Discrimination in Audit and Correspondence Studies. Journal of Human Resources 47(4), 1128–1157.

[59] Neumark, D. (2016). Experimental Research on Labor Market Discrimination. NBER Working Paper No. 22022.

[60] Neumark, D., I. Burn and P. Button (2015). Is It Harder for Older Workers To Find Jobs? New and Improved Evidence from a Field Experiment. NBER Working Paper No. 21669.

[61] Neumark, D., I. Burn and P. Button (2016). Experimental Age Discrimination Evidence and the Heckman Critique. American Economic Review, 106(5), 303–08.

[62] Neumark, D., R.J. Bank and K.D. Van Nort (1996). Sex Discrimination in Hiring in the Restaurant Industry: An Audit Study. Quarterly Journal of Economics 111(3), 915–42.

[63] Neumark, D. and J. Rich (2016). Do Field Experiments on Labor and Housing Markets Overstate Discrimination? Re-Examination of the Evidence. NBER Working Paper No. 22278.

[64] Olian, J.D., D.P. Schwab and Y. Haberfeld (1988). The Impact of Applicant Gender Compared to Qualifications on Hiring Recommendations: A Meta-Analysis of Experimental Studies. Organizational Behavior and Human Decision Processes, 41(2), 180–195.

[65] Petit, P. (2007). The Effects of Age And Family Constraints on Gender Hiring Discrimination: A Field Experiment in the French Financial Sector. Labour Economics 14, 371–391.

[66] Phillips, C. (2016). Do Comparisons of Fictional Applicants Measure Discrimination When Search Externalities Are Present? Evidence from Existing Experiments. Working Paper.

[67] Riach, P. A. and J. Rich (1987). Testing for Sexual Discrimination in the Labor Market. Australian Economic Papers 26, 165–178.

[68] Riach, P. A. and J. Rich (2002). Field Experiments of Discrimination in the Market Place. Economic Journal 112, F480–F518.

[69] Riach, P. A. and J. Rich (2006). An Experimental Investigation of Sexual Discrimination in Hiring in the English Labor Market. Advances in Economic Analysis & Policy 6(2), article 1.

[70] Rosen, B. and T.H. Jerdee (1974). Influence of Sex Role Stereotypes on Personnel Decisions. Journal of Applied Psychology, 59(1), 9–14.

[71] Schein, V. E. (1973). The Relationship Between Sex Role Stereotypes and Requisite Management Characteristics. Journal of Applied Psychology 57(2), 95-100.

[72] Scherr, A., C. Janz and S. Müller (2013). Readiness to Discriminate in Vocational Training: Results and Implications of a Survey of Firms. Soziale Probleme, 24, 245–270.

[73] Sharp, C. and R. Post (1980). Evaluation of Male and Female Applicants for Sex-Congruent and Sex-Incongruent Jobs. Sex Roles 6 (3), 391–401.

[74] Statistisches Bundesamt (Destatis) (2016). Verdienststrukturerhebung 2014 Fachserie 16 Heft 1.

[75] Stephan, G., M. Dütsch, C. Gückelhorn and O. Struck (2014). When are Bonus Payments for Managers Perceived as Fair? Results from a Quasi-Experiment. Economics Letters, 125 (1), 130–133.

[76] Upright, C. (2017). The Converging Gender Wage Gap 1980-2012. Contexts, 16(1), 72–74.

[77] Weichselbaumer, D. (2004). Is it Sex or Personality? The Impact of Sex-Stereotypes on Discrimination in Applicant Selection. Eastern Economic Journal 30, 159–86.

[78] Wooldridge, J.M. (2001). Econometric Analysis of Cross Section and Panel Data. MIT Press Books, The MIT Press, Edition 1, Volume 1, Number 0262232197, January.

[79] Zhou X., J. Zhang and X. Song (2013). Gender Discrimination in Hiring: Evidence from 19,130 Resumes in China. http://dx.doi.org/10.2139/ssrn.2195840.

A Appendix

A.1 The unobserved variance critique in vignette studies

In section 2 we argue that the estimate of discrimination obtained in a vignette study is robust to the unobserved variance critique under certain assumptions. In this appendix, we formalize the argument. To this end, we follow the model and reasoning of Neumark (2012) and Heckman (1998).

Denote gender by $g \in \{0, 1\}$, the firm by f, and suppose that the individual productivity X of an applicant consists of two components, namely observable productivity characteristics X_o and unobservable productivity characteristics X_u such that $X = (X_o, X_u)$. The productivity P of an applicant depends on his individual productivity characteristics and the requirements of the firm or, more generally, the firm characteristics such that $P = P(X, f, g) = P(X, f) = X_o\beta + X_u + f$, where the second equality sign follows if we assume that gender does not influence productivity and the third equality sign follows if we assume productivity to be linear in the different productivity components. Define the treatment T of a person of gender g and productivity P at firm f as $T(P(X, f), g)$. Define discrimination as $T(P(X, f), g = 1) \neq T(P(X, f), g = 0)$ or equivalently, if we assume that the treatment is additive in its arguments, $T(P(X, f), g) = P(X, f) + \gamma g$ with $\gamma \neq 0$.

In both our vignette study and correspondence studies researchers "standardize" on observable productivity characteristics, that is, they equalize the observable applicant characteristics, X_o, at a certain level between the applicants, such that $X_o = X_o^s$ for all applicants. In contrast, researchers have no control over the unobservable characteristics X_u. Heckman and Seligman's unobserved variance critique shows that a difference in the variance of the unobserved productivity characteristics, X_u, can bias the estimates of discrimination, even if the mean of X_u is the same across groups.[41] The issue arises because the outcome variable is not linear in productivity. That is, in correspondence studies it is plausible to assume that the decision rule, e.g., to offer an interview, depends on whether the perceived productivity P exceeds a cut-off such that, for $j \in \{0, 1\}$,

$$T(P(X, f), g = j) = \begin{cases} 1 & \text{if } P(X, f) \geq c \\ \\ 0 & \text{if } P(X, f) < c. \end{cases}$$

[41] The possibility that the mean of X_u differs between the groups under consideration is Heckman and Siegelman's first point of criticism. To identify discrimination, the equality of the average unobserved productivity-related factors has to be assumed.

Hence, if there is no discrimination,

$$T(P(X,f),g=j)=\begin{cases}1 & \text{if } X_o^s\beta+X_u^{g=j}+f\geq c\\ \\ 0 & \text{otherwise,}\end{cases}$$

for $j\in\{0,1\}$. Correspondence studies usually employ a probit model. Thus, let us assume that the unobservable productivity characteristics $X_u^{g=0}$ and $X_u^{g=1}$ are normally distributed and both have an equal conditional mean of zero, i.e., assume that $E(X_u^{g=1}|X_o^s)=E(X_u^{g=0}|X_o^s)=0$. Further assume that the unobservable productivity characteristics have different standard deviations, that is $\sigma_u^{g=1}\neq\sigma_u^{g=0}$. Then, denoting by Φ a standard normal distribution function

$$Pr(T(P(X,f),g=1)=1|X_o^s)=\Phi\left(\frac{X_o^s\beta-c}{\sigma_u^{g=1}}\right),$$

and

$$Pr(T(P(X,f),g=0)=1|X_o^s)=\Phi\left(\frac{X_o^s\beta-c}{\sigma_u^{g=0}}\right),$$

see, e.g., Wooldridge (2001).[42] Now, if $\sigma_u^{g=1}\neq\sigma_u^{g=0}$, $Pr(T(P(X,f),g=1)=1|X_o^s)\neq Pr(T(P(X,f),g=0)=1|X_o^s)$. Hence, even when there is no discrimination ($\gamma=0$), the outcomes differ. Therefore, discrimination is not identified.[43]

Now, suppose that the 10-point scale which employers use to make their evaluations does not restrict the evaluations of the respondents, but captures all evaluations that the respondents possibly want to make. In this case, we do not face a binary response model, and the treatment is linear in productivity:

$$T(P(X,f),g=j)=X_o^s\beta+X_u^{g=j}+f$$

for $j\in\{0,1\}$, where we again assume gender discrimination to be absent. Specifying the conditional mean, we have:

$$E(T(P(X,f),g=1)|X_o^s)=X_o^s\beta+f,$$

$$E(T(P(X,f),g=0)|X_o^s)=X_o^s\beta+f.$$

Thus, even if $\sigma_u^{g=1}\neq\sigma_u^{g=0}$, it holds that $E(T(P(X,f),g=1)|X_o^s)=E(T(P(X,f),g=0)|X_o^s)$, since differences in the variance of X_u do not matter for the identification. We are in the standard OLS framework and the crucial assumption for identification is the usual $E(X_u|X_o^s)=0$, which we have already assumed above.[44]

[42] By assuming that f is independent of $X_u^{g=j}$ and by assuming normality of f, we can for simplicity ignore the firm characteristic.

[43] As shown in Neumark (2012) in this case a heteroskedastic probit model can be estimated to identify discrimination.

[44] If $\sigma_u^{g=1}\neq\sigma_u^{g=0}$, however, standard errors need to be heteroskedasticity-adjusted.

Note that this analysis only holds true as long as we abstract from the discrete nature of our 10-point outcome variable. Otherwise, an ordered probit model is appropriate for estimating gender discrimination. In this case, heteroskedasticity of the errors changes the form of the response probabilities and discrimination is not identified. The same holds true if the 10-point scale restricts the evaluations of the respondents. We can treat this case as the data being censored. We cover both cases by estimating a heteroskedastic ordered probit model as a robustness check (see Section 4.2).

A.2 Gender difference in evaluations: ordered probit estimation

Table 5: Marginal effects of ordered probit estimation

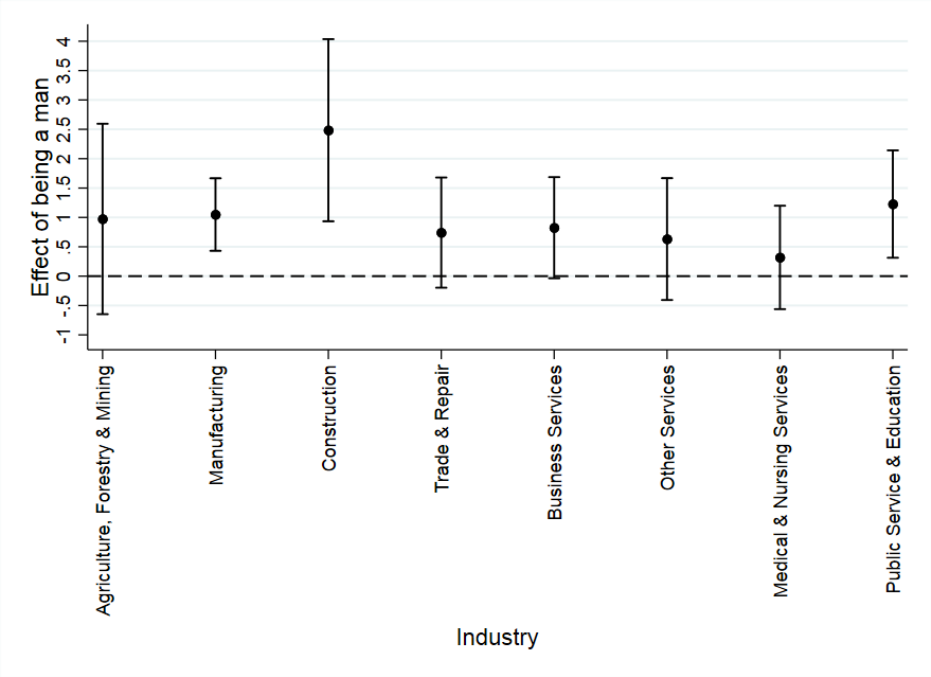
Evaluation	Marginal effect	Standard error
1	0.053*	0.029
2	0.019**	0.009
3	0.021**	0.010
4	0.013*	0.007
5	0.007*	0.004
6	-0.001	0.002
7	-0.010**	0.005
8	-0.025**	0.012
9	-0.027**	0.014
10	-0.050*	0.026
Observations	3164	

Notes: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. The second column reports the marginal effects of being female on the probability of observing an evaluation as indicated by the first column. The marginal effects are obtained from a random-effects ordered probit estimation of the evaluation on the vignette characteristics (see Table 3) and the vignette position using the sampling weights. Standard errors are obtained using the Delta-method and are clustered at the respondent-level.

A.3 Different industries

We can also group the firms into industries.[45] Our dataset contains firms from eight industries and we observe, on average, evaluations made by 79 respondents per industry.[46] Figure 6 shows the differences in evaluations by industries along with 95 percent confidence intervals obtained by regressing the evaluation on gender, industry dummies and the interactions between the industry dummies and gender (controlling for all other vignette characteristics and its position) using a RE regression. We find a difference that is significantly different from zero at the 5 percent level for "Manufacturing," "Construction," and "Public Service & Education," amounting to 1.05, 2.49, and 1.23, respectively. For the other industries, the coefficient of gender is not significantly different from zero at the 5 percent level, although women are evaluated worse than men in all industries.

Figure 6: Gender differences in evaluations by industry



Note: The figure presents the effect of being a man along with 95% confidence intervals obtained by a RE regression of the evaluation on a gender dummy, industry dummies, the interactions between the industry dummies and the gender dummy, all other vignette characteristics, and the vignette position using the sampling weight and clustering standard errors at the respondent level.

[45]The classification is based on the German Classification of Economic Activities, 2008 (WZ 08).
[46]Observations per category range from 183 (in the construction industry) to 754 (in manufacturing).